ESSAYS ON NONPARAMETRIC STRUCTURAL ECONOMETRICS: THEORY AND APPLICATIONS

BY ZHUTONG GU

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Economics

Written under the direction of

Roger W. Klein

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2017

ABSTRACT OF THE DISSERTATION

Essays on Nonparametric Structural Econometrics: Theory and Applications

by ZHUTONG GU

Dissertation Director: Roger W. Klein

My dissertation contains three papers in the theory and applications of nonparametric structural econometrics. In chapter 1, I propose a nonparametric test for additive separability of unobservables of unrestricted dimensions with average structural functions. Chapter 2 considers identification and estimation of fully nonparametric production functions and empirically tests for the Hicks-neutral productivity shocks, a direct application of the test proposed in chapter 1. In chapter 3, my authors and I study the semiparametric ordered response models with correlated unobserved thresholds and investigate the issue of corporate bond rating biases due to the sharing of common investors between bond-issuing firms and credit rating agencies. Brief abstracts are presented in order below.

Additive separability between observables and unobservables is one of the essential properties in structural modeling of heterogeneity in the presence of endogeneity. In this chapter, I propose an easy-to-compute test based on empirical quantile mean differences between the average structural functions (ASFs) generated by nonparametric nonseparable and separable models with unrestricted heterogeneity. Given identification, I establish conditions under which structural additivity can be linked to the equality of ASFs derived from the two commonly employed competing specifications. I estimate the reduced form regressions by Nadaraya-Watson estimators and control for the asymptotic bias. I show

that the asymptotic test statistic follows a central χ^2 distribution under the null hypothesis and has power against a sequence of $\sqrt{N}$-local alternatives. The proposed test statistic works well in a series of finite sample simulations with analytic variances, alleviating the computational burden often involved in bootstrapped inferences. I also show that the test can be straightforwardly extended to semiparametric models, panel data and triangular simultaneous equations frameworks.

Hicks-neutral technology implies the substitution pattern of labor and capital in a production function is not affected by technological shocks, first put forth by John Hicks in 1932. In this chapter, I consider the identification and estimation of fully nonparametric firm-level production functions and empirically test the Hicks-neutral productivity in the U.S. manufacturing industry during the period from 1990 to 2011. Firstly, I extend the proxy variable approach to fully nonparametric settings and propose a robust estimator of average output elasticities in non-Hick-neutral scenarios. Secondly, I show that the Hicks-neutral restriction can be converted to the additive separability between inputs and unobservables in a monotonic transformed model for which the proposed testing procedure can be directly applied. It turns out that there is substantial heterogeneity in the nonparametric output elasticities over various counterfactual input amounts. I also find that there were periods in the 90s when the non-Hicks technological shocks occur which coincide with the mass adoption of computing technology. However, the productivity has thereafter become Hicks-neutral into the 2000s. Controlling for sector-specific effects mitigate the non-Hicks-neutrality to some extend.

Previous literature on bond rating indicates that credit rating agencies (CRAs) may assign favorable ratings to bond-issuing firms that have a closer relationship. This not only implies the existence of firm-specific unobserved heterogeneity in the rating criteria but also makes some bond/firm characteristics endogenous, which is confirmed by our empirical results. In this chapter, my coauthors and I propose a semiparametric two-step index and location estimator of ordered response models that explicitly incorporates endogenous regressors and correlated random thresholds. We apply our model in the application of assessing bond rating bias of credit rating agencies. Methodologically, we first show that the heterogeneous relative thresholds can be identified using conditional shift restrictions

in conjunction with the control variables for the firm-CRA *liaison*. Then, we illustrate the estimation strategy in a heuristic manner and derive the asymptotic properties of the suggested estimator. In the application, we find significant overrating bias through varying thresholds as the liaison strengthens and those biases display heterogeneous patterns with respect to rating categories.

Acknowledgements

The completion of this dissertation is made possible by all the support, inspiration and hardworking of many people. Amongst them, I am most grateful to my advisor, Roger W. Klein, who has been patiently, constantly and encouragingly providing guidance through the many years of my Ph.D. Not only did he introduce me to the world of econometrics, but also he exemplified as a role model for me. From him, I learned precision and rigor as an econometrician, inquisitiveness and objectiveness as a researcher, and more importantly, diligence and patience as an educator.

I am also indebted to Tom J. Prusa and Norman R. Swanson, not only for being supportive members of my dissertation and job search committee, but also for providing incisive suggestions to help me improve the quality of this dissertation. Many thanks go to Chan Shen for her constructive comments as the outside member. I also would like to express special thanks to Jan De Loecker for his excellent lecture in empirical industrial organization at Princeton University, where I got the idea of my second chapter. I would also like to thank John Landon-Lane, Xiye Yang, Yuan Liao, Bruce Mizrach and Hilary Sigman for their invaluable support during my graduate study.

I want to dedicate this dissertation to all my family and friends who have kept me accompanied all these years, believed in me from the beginning and stayed by my side during arduous times. In particular, I want to express my gratitude to Shuyang Yang and Yixiao Jiang for our co-authorship and friendship. Shuyang encouraged me with her brilliant ideas and positive mindsets; Yixiao inspired me with his hardworking and persistence. I owe my warmest gratitude to a very inclusive group of people, "Pot Love", with whom I had most of the happiest moments. Finally, very special thanks go to my friends, Han Liu, Mingmian Cheng, Wukuang Cun, especially Long Feng, for their company over the many years at Rutgers University.

Dedication

To my family and friends

Table of Contents

List of Tables

List of Figures

Chapter 1

Additive Separability and Excess Unobserved Heterogeneity: A Nonparametric Test with Average Structural Functions

1.1 Introduction

This paper proposes a simple test for structural separability between observed regressors and unobserved heterogeneity with unrestricted dimensions in the presence of endogeneity. Nonseparability has important implications in structural economic modeling. On the one hand, economic theory rarely specifies that the unobservables enter the structural equations in an additive way and quite often, they do so in a nonlinear fashion. Furthermore, the restriction of additivity is equivalent to the absence of unobserved individual heterogeneity of marginal and treatment effects. On the other hand, from a modeling perspective, structural separability might be of testing interest in its own right. And nonseparability, sometimes, can be a key assumption to rationalize endogeneity. However, a slew of empirical research has only focused on additive models until recently. Roughly speaking, a large class of estimators such as 2SLS/IV estimators or within/differencing estimators in panel data, requires the assumption of structural separability. Whereas inconsistent estimates might be produced once the validity of additive unobservables is in question. Therefore, a test for structural additivity can be very useful in empirical microeconomic contents. However, a challenge comes from the fact that unobservables are often multi-dimensional. In most cases, even the dimension is unknown. The implication of the multiplicity is that structural functions cannot be identified without imposing significant shape restrictions or distributional assumptions. To circumvent this difficulty, I build our test on average structural functions (ASFs), which are identified via the control function approach. In the paper, I derive testable implications of structural additivity on the equivalence of ASFs generated by competing specifications. I argue that testing the latter should be more

appealing, since the former hypothesis is not testable in the presence of unobservables of unrestricted dimension. An easy-to-compute test statistic is proposed by combining information from empirical quantile mean (EQM) differences. The asymptotic properties of the test statistic are derived following U-statistic theorems of various orders. I illustrate the performance of the proposed test in a series of Monte Carlo simulations.

This paper contributes to the literature of nonparametric idenfication of nonseparable models and testing of additive separability. Despite its empirical importance, additive separability has only received limited attention. Lu and White [86] show that structural additivity can be transformed into a conditional independence assumption using a control function approach. They require either polynomial parametric structures or scalar monotonicity on an unobservable to establish the equivalence of tests. Su et al. [115] 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. This significantly restricts 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 [56] who propose a test in the context of sample selection, Heckman et al. [45] who consider testing for the correlated random coefficient models. Hoderlein and Mammen [47] 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. [114] in cross-sectional context and Hoderlein et al. [49] for panel data models. Lewbel et al. [81] 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. [45], Fan and Li [34], Schennach et al. [108], Sperlich et al. [113],etc.

More importantly, a notable fact is that heterogeneity in microeconomics is rarely unit-dimensional and quite often even the number of dimensions is not even known *a priori*. Unobservables often represents unobserved heterogeneity of consumer tastes, product attributes, productivity shocks, measurement errors, etc. This paper tries to clarify the benefits and costs of allowing fully flexible unobserved heterogeneity in the context of testing

additive separability. As argued in Browning and Carro [19], most empirical models permit less heterogeneity than is actually present. However, if shape restrictions or distributional assumptions are prohibited, there would not exist any equivalent tests for separability, a point made clearly in the next section. This is because the structural function cannot be identified in the presence of excess unobserved heterogeneity. But the ASF is identified, which could serve the purpose of specification testing. The identification of ASFs relies on the control function literature [16, 59, 60, 35, 65, 90, 32, 116, etc.]. In this paper, I try to fill this gap and propose an easy-to-implement test by relating structural separability to the equality between ASF generated by the two competing models. The testing implication is that when the true model is separable, ASF under both models are consistent; but only the former is consistent if $\mathbb{H}_0$ is not true. For illustrative purpose, I give several examples where the test has zero power. Furthermore, as motivated by Blundell and Powell [15], ASF should be the central object of estimation interest, since it suffices to answer many economic questions that empirical researchers care about. The test proposed here can be informative in terms of the consistency and efficiency of ASF estimators. In particular, by employing the additive error structure, a more efficient estimator of ASF is made available given the non-rejection of the null hypothesis.

The secondary contribution is that a nonparametric empirical quantile mean (EQM) test has been developed, in the spirit of Klein [72]. The idea of the test is to compare average differences between two functions in quantile regions of observables. For each quantile, it can be shown that the average will converge to a normal distribution in large samples. Combining information of all quantiles, one can straightforwardly obtain a Wald-type test statistic which has local power against root-N alternatives. The benefits of the EQM are mainly two-fold. First, compared with Kolmogorov-Smirnov or Cramer-von Mises test statistics, the nonparametric functions are only evaluated at sample points which eliminates the need to select fixed evaluation points, especially when the dimension of X is large. Second, it permits a closer investigation of the heterogeneous functions at each quantile region, facilitating the discovery of the anomalies in the data. It is also informative on where the power of the test is coming from, specifically which region of the sample is more likely to be rejected than others. In the limit, I show that it converges to a central χ^2

distribution under the null hypothesis, using the U-statistic theorems, and has asymptotic power against a sequence of $\sqrt{N}$-local alternatives. To control for the asymptotic bias, I employ the recursive bias correction technique recently developed by Shen and Klein [110]. 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 even moderate sample sizes, in terms of the empirical power and size calculations. This advantage translates into a substantial decrease in computing time, especially for large samples.

Finally, three extensions are presented to cover more empirical scenarios. First, I address the problematic "curse of dimensionality" by imposing parametric index restrictions. Despite all attractive properties that nonparametric testing entails, in real applications, semiparametric index models are often employed due to the high dimensionality of the control covariates. I accommodate the test in a two-step semiparametric scenario. While the finite dimensional parameters are estimated in the first step by Weighted Semiparametric Least Squares (WSLS), then the same testing procedure can immediately be applied on the estimated single index [106, 58, 57, 74, etc.]. In the second extension, once panel data are available, the cross-sectional procedure can be modified to test additivity of time-invariant unobserved individual-heterogeneity. With repeated observations over time, control variables usually take the form of individual-specific summarized measures, e.g. average value over time, provided that the exchangability condition holds [7, 94, etc.]. In the last extension, I consider the nonparametric nonseparable triangular simultaneous equations models. Following Imbens and Newey [60], the marginal cumulative distribution function (CDF) of the first stage error suffices to work as the control variable. The asymptotic null distribution of the test statistic is thus modified to take into account the variability of estimation of the "generated" variable in the first stage.

The rest of the paper is structured as follows. Section 1.2 provides motivations for testing additive separability and discusses identification issues of nonseparable models with excess heterogeneity. Section 1.3 reviews the identification results of ASFs under competing specifications and clarifies the relationship to testing additive structures. Section 1.4 provides the nonparametric test statistics in a heuristic manner. The asymptotic results are stated in Section 1.5. Next, finite sample performance is summarized in Section 1.6.

Extensions to semiparametric models, panel data and triangular simultaneous equation frameworks are presented in Section 1.7. Section 1.8 concludes the paper and discusses its limitations. All proofs are relegated in the appendix.

1.2 Nonseparability and Unobserved Heterogeneity

1.2.1 Nonseparability

Nonparametric nonseparable models have been gaining popularity in theoretical econometric works for the past decades. The single equation nonseparable model considered in this paper, as seen in Eq. (1.1), allows arbitrary interactions between observed and unobserved covariates, e.g. X versus ε.

$$Y = m(X, \varepsilon) \tag{1.1}$$

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 structural parameters of economic interest. For instance, model (1.1) can represent a nonparametric 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 unobservables pertaining to demand and cost conditions. Model (1.1) is also general enough to include an entire class of random coefficient models that are widely used in empirically modeling unobserved individual heterogeneity, e.g. $Y = X'\varepsilon$, where X and ε are conformable vectors.

Being a special case of model (1.1), a competing class of specifications disproportionately favored in empirical work, assumes the additive separable structure in which the unobservables can be collectively written as an added term,

$$Y = m_1(X) + m_2(\varepsilon) \tag{1.2}$$

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) [98, 22, 28, 97, 52, etc.] or control function approach [100, 99, 104, etc.]. If panel data are available, within and differencing estimators can apply to address the endogeneity arising from the correlation between individual heterogeneity and time-varying covariates [8, 14, etc.].

Given the above nested specifications, I define the following hypotheses that are of interest to empirical researchers.

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

The motivations for testing hypotheses $\mathbb{H}_0^*$ against $\mathbb{H}_1^*$ are mainly fourfold. 1). This is a test on the absence of unobserved individual heterogeneity in structural functions. Once $\mathbb{H}_0^*$ is rejected, 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 the example of estimating log of production functions, the additivity of errors amounts to the Hicks-neutral technology—output elasticities are not affected by productivity shocks.[1] 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 [41] show that the conditional mean restriction, often assumed in IV methods, only has identification power when the model is additive in unobservables. Schennach et al. [108] show that interpretation of the local indirect least squares (LILS) estimator is meaningful only under the separability of the structural equation that determines X. 3). There are more efficient estimators given the additional parametric structure under $\mathbb{H}_0^*$. Hahn [40] and Imbens and Wooldridge [61] note that the asymptotic variance bounds of average treatment effect (ATE) can be made

[1] See Gu [38] on the detailed discussion of the implications of Hicks-neutral technologies.

much smaller once additive separable unobservables can be validated. 4). In some structural models, testing separability could yield implications on testing endogeneity, a point taken from Imbens [59] and Imbens and Newey [60]. 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 the cost function for which $C_x > 0$ and $C_{xx} > 0$. 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. This can be undesirable since not many models would treat input choices as purely exogenous.

1.2.2 Non-identification under Excess Unobserved Heterogeneity

Before outlining the testing framework, I want to highlight the identification problem associated with multi-dimensional unobservables in nonseparable structural models. More importantly, I want to argue that the original hypothesis is not testable to the extent that unobserved heterogeneity is allowed to be modeled as flexibly as possible. For this reason, below I will modify the hypotheses of testing interest.

I begin by discussing a fallacy pointed out by Benkard and Berry [12]. They revisit the identification results of simultaneous equations models from Brown [18] and Roehrig [107] and show that a supporting lemma (called derivative condition) is incorrect. Although they consider the identification of certain features in the framework of simultaneous equations models, the following lemma could still shed light on the identification issue of the single-equation structural function considered in this paper.

Lemma 1.1 (Derivative Condition. Brown (1983, pp. 180-181)). *Let X and ε be independent random vectors and let $\widetilde{\varepsilon} = T(X, \varepsilon)$, where $T : \mathbb{R}^{d_X} \times \mathbb{R}^{d_E} \to \mathbb{R}^{d_E}$ is everywhere*

differentiable. Then $\widetilde{\varepsilon}$ is independent of X if and only if $\partial T(x,e)/\partial x = 0$ everywhere.

One direction is true that if the derivative is everywhere 0, then it indicates $T(\cdot,\cdot)$ is a degenerate function of X and independence holds trivially. The questionable one is the other direction unless the unobservable is univariate. This point is made clear from the simple example given below in the spirit of Benkard and Berry.

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

$$Y = \frac{X}{\sqrt{X^2+1}}\varepsilon_1 + \frac{1}{\sqrt{X^2+1}}\varepsilon_2$$

and

$$Y = \varepsilon_3$$

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. Formally, I follow the definition in Roehrig [107] and let $F_{X,\varepsilon}$ be the distribution and in conjunction with the structure S be a pair $(F_{X,\varepsilon}, \theta)$ that define the data generating process, where θ is the vector of finite or infinite dimensional parameters.

Definition 1.1. *Let F and F' be the distribution functions of (X,Y) implied by the structure S and S', then S and S' are observationally equivalent if $F = F'$.*

Definition 1.2. *The structure S is identified if there is no other S' that is observationally equivalent to S.*

The example above indicates that the structural function itself is not identified without further restrictions. As a consequence, it implies that our original hypothesis of $\mathbb{H}_0^*$ versus $\mathbb{H}_1^*$ is not testable in general because both nonseparable and separable models can deliver the same underlying data generating process, meaning they are observationally equivalent.

One solution is to impose additional structures to achieve identification [see 91]. 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. [115] 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 [86] transform the original hypothesis into a conditional independence condition. However, they lose equivalence unless $m_1(\cdot)$ is some polynomial function or scalar monotonicity in unobservables holds.

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 hardly find sound theoretical or empirical support. Another direction is to determine what can be identified without compromising the dimensionality of heterogeneity [16, 60, etc]. In many cases, 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 exactly what this paper is trying to do. 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 unosbervables, there are papers dealing with identification and estimation via restrictions like single index property [12, 92, 93, 25, 26, etc.], which may be exploited to develop other testing procedures.

1.3 Testing with Average Structural Functions

In this section, I first review the existing results on the identification of ASFs and then derive testable implications for structural separability. Define the ASF at $X = x$ of nonseparable models (1.3) as

$$ASF(x) \equiv g(x) = \int_{\mathcal{E}} m(x, e) dF_{\varepsilon}(e), \quad \forall x \in \mathcal{X} \tag{1.3}$$

where calligraphic letters denote the support on which F_{ε}, the CDF of ε admitting the Radon-Nikodym derivative, 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 [15], ASFs should be the central object of estimation interest and with which many important structural objects can be easily constructed. For instance, the average treatment effect (ATE) can be obtained from ASFs for when X is binary, like in Eq. (1.4),

$$ATE = g(1) - g(0) = \int_{\mathcal{E}} [m(1,e) - m(0,e)] dF_{\varepsilon}(e) \tag{1.4}$$

For continuous treatments, the average marginal effect (AME) in Eq. (1.5) is readily available provided the existence of $m_x(,e), \forall e \in \mathcal{E}$:

$$AME(x) = g'(x) = \int_{\mathcal{E}} m_x(x,e) dF_{\varepsilon}(e), \quad \forall x \in \mathcal{X} \tag{1.5}$$

If $Y = m(X, \varepsilon)$ denotes a nonparametric production function, then the AME measuring average marginal products conditional on input choices, can be exploited to calculate average output elasticities and return-to-scale. The property of manipulating input choices while holding the distribution of unobserved productivity (and other unobservables) fixed would be attractive to policymakers, industry specialists and firm managers. A related object is the average derivative, $E[\partial g(X, \varepsilon)/\partial x]$, which summarizes the marginal effect of X on Y over the whole population. Given that the AME is identified at each point in the support, the average derivative can be simply recovered by taking the expectation over X.[2] Imbens and Newey [60] study the identification of ASF as well as a number of structural parameters of economic interest.[3]

Now recall the nonseparable model (1.1), where the unknown structural function is defined on $\mathcal{X} \times \mathcal{E}$ and $\mathcal{X} \subset \mathbb{R}^{d_X}$ exhibiting continuous variation and $\varepsilon \subset \mathbb{R}^{\infty}$, referring excess unobserved heterogeneity.[4] The identification of ASF without endogeneity is trivial, as suggested by the reduced form regression $g(x) = E(Y|X = x), \forall x \in \mathcal{X}$. Unfortunately,

[2] Average structural derivative over some region, $\mathcal{X}^0$, could be identified if ASF is only defined on $\mathcal{X}^0$.

[3] Another meaning measure they consider is the quantile structural function, i.e. $QSF(\tau, x) = q(\tau, x), \quad \forall x \in \mathcal{X}$ where $q^{-1}(y, x) \equiv \int_{\mathcal{E}} \mathbf{1}(m(x,e) < y) dF_{\varepsilon}(e) = \Pr(m(x, \varepsilon) < y) = \tau$.

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

many economic models would include at least one endogenous regressor. For instance, the "simultaneity bias" could arise from the dependency of choices of input on the unobserved productivity shocks when estimating firm or plant 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 [60]. 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 several ways to obtain the control variates, V, in empirical contents. 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. A famous example is the production function estimation where a "proxy" variable can be backed out from the investment functions to control for the unobserved productivity shocks. I will come back and elucidate these issues in Section 1.7. For now, I just presume the control variables V satisfying Assumption I-1 and I-2 to be 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 implication.

1.3.1 Identification of ASF of Nonseparable Models

Proposition 1.1 is borrowed from Blundell and Powell [16], Imbens and Newey [60], etc. It shows 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 1.1. *Under Assumption I.1 and I.2, $g(\cdot)$ defined in Eq. (1.3) is identified at each $x \in \mathcal{X}$,*

$$g(x) = \int_{\mathcal{V}} C(x, v) dF_V(v) \tag{1.6}$$

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 1.1, related "structural" parameters, provided existence, are subsequently obtained. In the literature of program evaluation, one of the most important parameters, ATE in Eq. (1.4), can be identified in Eq. (1.7),

$$ATE = \int_{\mathcal{V}} [C(1, v) - C(0, v)] dF_V(v) \tag{1.7}$$

If X is a continuous treatment, $AME(x)$ in Eq. (1.5) is identified in Eq. (1.8), provided that the partial derivatives of $C(\cdot, \cdot)$ exist.

$$AME(x) = \int_{\mathcal{V}} C_x(x, v) dF_V(v) \tag{1.8}$$

where $C_x(\cdot, \cdot)$ denotes the partial derivative with respect to X. More generally, some policy changes may involve transformation of structural functions. Suppose $\tau(\cdot)$ is any linear functional operator on $m(x, e)$ with finite first moment of $\tau(m(x, e)), \forall (x, e) \in (\mathcal{X} \times \mathcal{E})$. The

transformed ASF is identified in Eq. (1.9),

$$\int_{\mathcal{E}} \tau(m(x,e))dF_{\varepsilon}(e) = \int_{\mathcal{V}} \tau(C(x,v))dF_{V}(v) = \tau(g(x)) \tag{1.9}$$

Note that the linear functionals include, but are not limited to, the weighted averages of structural functions and AME.

1.3.2 Identification of ASF of Separable Models

A popular subclass of models incorporates the structure that observables and unobservables are additive as in Eq. (1.2). Such models impose substantial restrictions on the way how unobserved heterogeneity enters. It further indicates constant partial effects given observed covariates and rules out an entire class of models with correlated random coefficients. Despite the reduced generality, it has received most attention in both theoretical and empirical works. Admittedly, the ASF of model (1.2) is immediately identified under Assumption I.1 and I.2 through Proposition 1.1 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. [99].

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

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

$$C(x,v) = m_1(x) + h(v), \forall (x,v) \in \mathcal{X} \times \mathcal{V} \tag{1.10}$$

Assumption I.2' Nonexistence of additive functional dependence. $\Pr(\delta(X)+\gamma(V) = 0) = 1$ implies there is a constant c_δ that $\Pr(\delta(X) = c_\delta) = 1$, 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. Formal identification results are given in Proposition 1.2.

Proposition 1.2. *Under Assumption I.1' and I.2' (or Assumption I.1 and I.2), a). $m_1(\cdot)$ and $h(\cdot)$ in Eq. (1.10) is identified up to an additive constant for each $(x,v) \in \mathcal{X} \times \mathcal{V}$. b). $g(\cdot)$ defined in Eq. (1.3) is identified at each $x \in \mathcal{X}$,*

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

Without loss of generality, one can normalize that $E[h(V)] = c_h = 0$, essentially attributing all constant terms 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. (1.11).[5]

$$h(v) = \int_{\mathcal{X}} C(x,v)dF_X(x) - E(Y) \tag{1.11}$$

The additive structure of model (1.1) provides us with more information which can be exploited in recovering the ASF through the one-step backfitting procedure in Linton [84]. 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. (1.12),

$$a(x) = E(Y - h(V)|X = x), \ \forall x \in \mathcal{X} \tag{1.12}$$

In Proposition 1.3, 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.

Proposition 1.3. *Under Assumption I.1 and I.2, for each $x \in \mathcal{X}$, a). for additive models of (1.2), $a(x) = g(x)$; b). for nonseparable models of (1.1), $a(x) = g(x)$ if and only if the*

[5] 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.

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

The equality in Proposition 1.3 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, below I present 3 examples of nonseparable models that can produce the equality. Example 1 is to illustrate that a nonseparable model can be written 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 a non-additive CEF, the ASF of a nonseparable model might 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, so no endogeneity arises. 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 1.3 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 [86] who argue that testing separability according to CEF has no power in the example 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 1.3 does hold in all above examples.

1.3.3 Testing Implications

As discussed in the introduction, this paper makes the very first attempt to test additive separability with unrestricted unobservables. Now recall the hypotheses of testing interest outlined previously,

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

Unfortunately as discussed in Section 1.2, no consistent test exists against global alternatives due to the non-identification of structural functions once excess heterogeneity is allowed. Hence in this paper, instead of testing $\mathbb{H}_0^*$ against $\mathbb{H}_1^*$, I consider a more interesting 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 two competing specifications are identical. The power of this test comes from the fact that $g(\cdot)$ in Eq. (1.6) recovers the ASF for both models whereas $a(\cdot)$ in Eq. (1.12) only recovers the ASF for additive models (and a small class of nonseparable models satisfying the condition stated in Proposition 1.3). 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. Second, the test still has reasonable power against additive separability, though not against global alternatives for $\mathbb{H}_0^*$, as can be seen from our finite sample simulations in Section 1.6. Finally, were ASFs and its variants sufficient to answer the research questions, there would be no need to test the original 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.

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. So researchers are advised to be mindful 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 hypothesis—$\mathbb{H}_0$ versus $\mathbb{H}_1$.

1.4 Estimation and Testing

1.4.1 Estimation

I first discuss the nonparametric estimator for the CEF which is the central building block for the test statistic. In this paper, I focus on the Nadaraya-Watson (or local constant) estimator [95] for conditional mean functions. Other nonparametric smoothers such as local polynomials and sieve estimators can be applied as well. To facilitate the proof of asymptotic theory, the 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 $C(x,v) = E(Y|X=x, V=v)$. Given any non-boundary set of points, $(x,v) \in \mathcal{X} \times \mathcal{V}$, the preliminary kernel estimator is defined in Eq. (1.13),

$$\widehat{C}_0(x,v) = \frac{\sum_{i=1}^N K_{h_1}(X_i - x)K_{h_1}(V_i - v)Y_i}{\sum_{i=1}^N K_{h_1}(X_i - x)K_{h_1}(V_i - v)} \tag{1.13}$$

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

To make sure that the asymptotic bias vanishes faster than $\sqrt{N}$, I suggest the recursive nonparametric conditional mean estimator recently proposed by Shen and Klein [110], due to its bias-reducing property.[6] Simply put, I firstly construct the local bias from the preliminary kernel estimator, e.g. $\widehat{\delta}_i(x,v) = \widehat{C}_0(X_i, V_i) - \widehat{C}_0(x,v)$ and then apply the kernel estimator again on the "bias-free" dependent variable, e.g. $Y_i - \widehat{\delta}_i(x,v)$. So the bias-reducing conditional mean estimator is thus obtained in Eq. (1.14).

$$\widehat{C}(X_l, V_j) = \frac{\sum_{i\neq j,l}^N K_{h_1}(X_i - X_l)K_{h_1}(V_i - V_j)[Y_i - \widehat{\delta}_i(X_l, V_j)]}{\sum_{i\neq j,l}^N K_{h_1}(X_i - X_l)K_{h_1}(V_i - V_j)} \tag{1.14}$$

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

Next I consider the estimation of ASFs. Linton and Nielsen [83] suggest a marginal integration method while Newey [96] 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. 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. (1.15),

$$\widehat{g}(X_l) = \frac{1}{N-1}\sum_{j\neq l}^N \widehat{C}(X_l, V_j), \quad \forall l = 1, \cdots, N \tag{1.15}$$

[6] Other bias reducing methods such as higher order kernels, local smoothing should work in theory. However, it is found that using higher order kernels are likely to produce unreasonable results in the finite sample simulations.

Likewise, $h(\cdot)$ defined in Eq. (1.11) can be estimated in the similar fashion in Eq. (34)

$$\widehat{h}(V_j) = \frac{1}{N-1}\sum_{i\neq j}^{N}\widehat{C}(X_i, V_j) - N^{-1}\sum_{i=1}^{N} Y_i, \quad \forall j = 1, \cdots, N \tag{1.16}$$

where the mean of Y subtracted resembles the sample analog of the unconditional expectation, $E(Y)$. This subtraction ensures the normalization of recentering such that the unconditional mean of $h(\cdot)$ is now 0.[7]

Now consider the ASF estimator, $\widehat{a}(\cdot)$ of the "additive" model. I borrow the idea from Linton [84] 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 [44] and is thought to yield the most efficient estimator since it finds the unique orthogonal projection of Y onto the space of additive functions providing the best mean square error approximation. However, such iterative nature not only requires intensive computational effort but cannot guarantee convergence sometimes. 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. (1.17).

$$\widetilde{a}(X_l) = \widehat{E}_{h_2}(Y_i - h(V_i)|X_l), \quad l = 1, \cdots, N \tag{1.17}$$

where $\widehat{E}$ is still the bias-reducing recursive conditional mean estimator similar to Eq. (1.14), with the bandwidth $h_2 \to 0$ as $N \to \infty$. By simply substituting $\widehat{h}(V_i)$ for the unknown

[7] Note that the CEF estimator in constructing $\widehat{h}(\cdot)$ could be potentially different from the one in Eq. (1.15) in terms of bandwidth and kernel choices.

function $h(V_i)$, one can obtain the feasible estimator $\widehat{a}(\cdot)$ in Equation (1.18),

$$\widehat{a}(X_l) = \widehat{E}_{h_2}(Y_i - \widehat{h}(V_i)|X_l) = \frac{\sum_{i\neq l}^{N} K_{h_2}(X_i - X_l)[Y_i - \widehat{h}(V_i) - \widehat{\delta}_i(X_l)]}{\sum_{i\neq l}^{N} K_{h_2}(X_i - X_l)}, \quad l = 1, \cdots, N \tag{1.18}$$

where $\widehat{\delta}_i(X_l)$ is the bias from preliminary estimators defined in the similar way as $\widehat{\delta}_i(X_l.V_j)$. Our estimator differs from Linton [84] 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 utilized.

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 the ASF when $\mathbb{H}_0$ is true as can be seen from the finite sample results.[8]

1.4.2 Test Statistics

The specification test of $\mathbb{H}_0$ falls into the class of testing on the distance between two functions being uniformly 0. To this end, the Kolmogorov-Smirnov or Cramer-von Mises test statistics are often applied. But in this paper, I adopt a simpler test that combines information from empirical quantile mean (EQM) differences. The test idea is firstly mentioned in Klein [72] in testing parametric error distributions versus semiparametric binary choice models.

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

$$D(X_i) \equiv g(X_i) - a(X_i), \quad \forall i, \ldots, N \tag{1.19}$$

Under $\mathbb{H}_0$, $D(X_i) = 0$ for each i almost surely. To proceed, divide the whole sample into P_N

[8] However, our ASF estimator under $\mathbb{H}_0$ is not the most efficient estimator. For efficiency, see Linton [85].

number of even subsamples or quantile regions thereafter, within 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. However, the choices of P_N might have some implications for the power of the test and I postpone the discussion of this until the finite sample simulations.

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

$$T_N^p \equiv N^{-1} \sum_{i=1}^{N} t_i^p D(X_i), \quad p = 1, \ldots, P_N \tag{1.20}$$

where the quantile-trimming indicator is defined in Eq. (1.21),

$$t_i^p \equiv \mathbf{1}\left\{\min[c_{lb}, q_X(p - 1/P_N)] \leqslant X_i < \max[q_X(p/P_N), c_{ub}]\right\} \tag{1.21}$$

where $q_X(\cdot)$ is the quantile function of X, i.e. $q_X(\tau) = \inf\{x : F_X(x) \geqslant \tau\}$. c_{lb}, c_{ub} are predetermined fixed lower and upper bounds, respectively, to ensure non-existence of significant 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 regular central limit theorem. A Wald-type statistic in Eq. (1.22) could be thus constructed,

$$W_N \equiv N T_N' \Omega^{-1} T_N \tag{1.22}$$

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 1.2 for explicit expressions.

One of the benefits of using EQM test is that empirical observations are evaluated to construct the test statistic W_N. So there is no need to select weighting functions on $\mathcal{X}$ or carry out numerical integration, especially when the dimension of X is large.

Moreover, dividing sample into subregions enables researchers to have a closer look across quantiles and is conducive to discover anomalies hidden in the data. Quite often, with a littler 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 characteristics.

A feasible test statistic is made possible by substituting unknown objects with corresponding estimators, like in Eq. (1.23)

$$\widehat{W}_N = N\widehat{T}_N'\widehat{\Omega}_N^{-1}\widehat{T}_N \tag{1.23}$$

where

$$\widehat{T}_N = (\widehat{T}_N^1, \ldots, \widehat{T}_N^P)' \tag{1.24}$$

and $\widehat{\Omega}$ is the consistent estimator of Ω given explicitly in Eq. (1.34) in the Corollary 1.2.1,

$$\widehat{T}_N^p = N^{-1}\sum_{i=1}^{N} \widehat{t}_i^p \widehat{D}(X_i), \quad p = 1, \ldots, P_N \tag{1.25}$$

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. (1.26).

$$\widehat{t}_i^p \equiv \mathbf{1}\left\{\min[c_{lb}, \widehat{q}_X(p - 1/P_N)] \leqslant X_i < \max[\widehat{q}_X(p/P_N), c_{ub}]\right\} \tag{1.26}$$

where the quantile function is defined in the following way,

$$\widehat{q}_X(\tau) = \inf\left\{x : N^{-1}\sum_{i=1}^{N} \mathbf{1}(X_i > x) \geqslant \tau\right\}$$

A final remark is concerning the choice of the 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. Instead, one is suggested to experiment with various values for robustness check, e.g. $P_N = 4, 6$ or 8 as in our Monte Carlo studies.

1.5 Asymptotic Properties

In this section, I first present the asymptotic theory of the nonparametric test statistic under the null hypothesis and then its power function against a sequence of local alternatives. Before stating the asymptotic assumptions, notations are simplified in the following way and will be carried through the rest of the paper.

Notation. 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 bounded away from 0. Also let $f^*(x, v) \equiv f_X(x) f_V(v)/f_U(u)$ for any $(x, v) \in \mathcal{U}_0$.

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). $\mathrm{Var}(Y|X, V) < \infty$. 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 with respect to Lebesgue measure. 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 $u, \widetilde{u} \in \mathcal{U}_0$, $|D^j E(Y|U = u) - D^j E(y|U = \widetilde{u})| \leqslant C||u - \widetilde{u}||$, 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). $P_N = o(N)$. iii). $d = d_X + d_V < 4$.

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. Nevertheless, the asymptotic theory developed here can be readily extended to weakly dependent time series. 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 kernels. In the following theory, only the second order kernel ($\nu = 2$), such as the standard normal, is required in conjunction with the recursive bias-reducing procedure.[9] Assumption A.4 guarantees the uniform consistency for the kernel estimator of conditional means. Assumption A.5 restricts the choices of bandwidth as well as number of quantile regions. 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 MSE. Assumption A.5 iii). restricts the dimension of observables to be less than 4 if optimal weights are used. It is possible to extend this restriction. Nevertheless, in empirical settings, number of control covariates are often of large dimension. I suggest a semiparametric version of the test and postpone the discussion in Section 1.7. Assumption A.6 states that the weighting matrix defined in Eq. (1.33) is invertible and is essentially the non-degeneracy condition of the test statistic.

In what follows, I show that the asymptotic null distribution of pth-quantile average difference in Theorem 1.1, with the scratch of the proof outlined below. All details and supporting lemmas are given in the appendix.

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

1.5.1 Asymptotic Null Distribution

Theorem 1.1. *Suppose $\mathbb{H}_0$ is true, under Assumption I.1, I.2 and A.1-A.6, for any $p \in (1, 2, \cdots, P_N)$, it is true that*

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

where

$$\Omega_p = E(\xi_i^p \xi_i^{p'}) \tag{1.27}$$

and the influence function, ξ_i^p, is defined in Eq. (1.28)

$$\xi_i^p \equiv [t_i^p + E(t^p|V_i)]f^*(X_i, V_i) - t_i^p]\epsilon_i + E(t^p)h(V_i) \tag{1.28}$$

where

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

Theorem 1.1 says the quantile average difference defined in Eq. (1.25) converging to a normal distribution at the parametric rate under $\mathbb{H}_0 : g(X_i) = a(X_i)$. The proof of the above theorem can be roughly divided in three steps . Firstly, I show that the estimated quantile (trimming) indicator can be replaced by its true counterpart plus reminder terms converging faster than $\sqrt{N}$. Secondly, I show that the empirical difference can be decomposed into three components through the substitution of 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 reminders and then the standard CLT applies. All the rest theorems and corollaries rely critically on Theorem 1.1.

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. (1.21) and Eq. (1.26). 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 .2 shows it is equal to $o_p(N^{-1/2})$ via the Cauchy-Schwartz inequality. 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 $\mathbb{H}_0$ almost surely.

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

Recall $\widehat{a}(\cdot)$ in Eq. (1.18) suffers from problems 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 kernel estimator of Δ given X_i as before.[10] 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. (1.29)

$$\widetilde{T}_N^p \equiv D_N^g + D_N^a + D_N^h \tag{1.29}$$

where by definition

$$D_N^g = N^{-1} \sum_{i=1}^{N} t_i^p (\widehat{g}(X_i) - g(X_i)) \tag{1.30}$$

$$D_N^a = -N^{-1} \sum_{i=1}^{N} t_i^p (\widetilde{a}(X_i) - a(X_i)) \tag{1.31}$$

$$D_N^h = N^{-1} \sum_{i=1}^{N} t_i^p \widehat{E}(\Delta(V_i) | X_i) \tag{1.32}$$

Step 3: By the U-statistic theorems of various orders, D_N^g, D_N^a and D_N^h can be

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

represented as sample means, characterized by Lemma .2, .3 and .5, respectively.[11] Therefore, we can rewrite $\widehat{T}_N^p$ as an influence function plus asymptotically negligible terms at $\sqrt{N}$-rate.

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

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 $h(\cdot)$. Then the standard CLT applies to the sample average while the remainder vanishes in the limit. To have a cleaner expression, I apply the recursive kernel estimator to ensure that the asymptotic biases vanish faster than $\sqrt{N}$ so that the vector of quantile mean differences will center at 0.

Theorem 1.2 below combines the information of the vector $\widehat{T}_N$ which, under the null, follows the asymptotic multivariate normal distribution with an invertible diagonal covariance matrix. The quantile test statistic then converges asymptotically to the χ_P^2 distribution with the degree of freedom equal to the predetermined number of quantile regions. In Corollary 1.2.1, I derive the asymptotic null distribution for the feasible test statistic by plugging-in a consistent covariance estimator of Ω.

Theorem 1.2 (The infeasible test statistic $\widehat{W}_N^0$)**.** *Suppose Assumption I.1, I.2 and A.1-A.6 hold, under $\mathbb{H}_0$, it follows that*

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

where $\widehat{W}_N^0 = N\widehat{T}_N'\Omega^{-1}\widehat{T}_N$, with $\widehat{T}_N$ in Eq. (1.24) and Ω in Eq. (1.33)

$$\Omega = E(\xi_i\xi_i') \tag{1.33}$$

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

Corollary 1.2.1 (The feasible test statistic $\widehat{W}_N$)**.** *Suppose Assumption I.1, I.2 and A.1-A.6*

[11] One technical simplification is to replace the estimated density denominator with the truth, guaranteed by the preliminary Lemma A 2.

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.* (1.24) *and* $\widehat{\Omega}_N$ *in Eq.* (1.34)*.*

The consistent estimator of covariance matrix $\widehat{\Omega}_N$ in Theorem 1.1 is therefore obtained in the following way,

$$\widehat{\Omega}_N = N^{-1}\sum_{i=1}^{N}\widehat{\xi}_i\widehat{\xi}'_i \tag{1.34}$$

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.

$$\widehat{\xi}_i^p \equiv \{[\widehat{t}_i^p + \widehat{E}(t^p|V_i)]\widehat{f}^*(X_i, V_i) - t_i^p]\}\hat{\epsilon}_i + \overline{t}^p\widehat{h}(V_i) \tag{1.35}$$

where the overhead bar represents the mean and

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

1.5.2 Local Alternative Analysis

Developing the global power function can be extremely difficult in the nonparametric testing contents, but one can study the local power property by considering a sequence of local alternatives:

$$\mathbb{H}_{1N} : g(X_i) - a(X_i) = N^{-1/2}r(X_i)$$

where r is a non-constant measurable function with $r_0 \equiv \lim_{N\to\infty} E[r(X)^2] < \infty$.

Theorem 1.3 (Asymptotic Local Power)**.** *Suppose Assumptions I.1, I.2 and A.1-A.6 hold, then under* $\mathbb{H}_{1N}$*,* $\widehat{T}_N \sim N(\widetilde{\boldsymbol{r}}_0, \Omega)$*, this implies* $W_N \sim \chi^2_P(\lambda)$*, where the noncentrality parameter* $\lambda = \sum_{p=1}^{P}\widetilde{r}_{0p}^2$ *and* $\widetilde{r}_{0p} = E[t_i^p r(X_i)]$*. Therefore, the asymptotic local power*

function is given by $\Pr(W_N > w|\mathbb{H}_{1N}) = 1 - Q_{P/2}(\sqrt{\lambda}, \sqrt{w})$ *with Marcum Q-function* $Q_M(a, b)$.

Theorem 1.3 implies that the test has non-trivial power against a sequence of local alternatives converging at the parametric $\sqrt{N}$-rate. The nonparametric test here attains the $\sqrt{N}$ rate because one additional averaging of nonparametric estimators is taken over all coordinates. This property is not shared by many other nonparametric tests. Theorem 1.4, given below, is a direct implication of Theorem 1.3 and it shows that our test is consistent under this scenario.

Theorem 1.4 (Consistency of the Test)**.** *Suppose Assumptions I.1, I.2 and A.1-A.6 hold, then* $\Pr(W_N > C_N|H_{1N}) = 1$ *as* $N \to \infty$ *for any* $C_N = o(N)$.

1.5.3 A Bootstrapped Version of the Test

For the sake of completeness, I also present a bootstrapped version of the test despite the fact that it may be very time-consuming for large dataset. In the following, I list the step-by-step procedure for computing bootstrapped empirical sizes and powers. $\mathbb{H}_0$, additive separability, need to be imposed to generate bootstrapped samples. As a result, one has an additive model like the following, which is true under $\mathbb{H}_0$,

$$Y_i = m_1(X_i) + h(V_i) + \epsilon_i$$

Note that the partial mean ASF estimator, $\widehat{g}(\cdot)$, always consistently estimates $m_1(\cdot)$ in any situation.

1. Obtain the preliminary estimates of $\{\epsilon_i\}_1^N$, i.e. $\hat{\epsilon}_i = Y_i - \widehat{C}(X_i, V_i)$.
2. Draw a bootstrapped sample $\{\epsilon_i^*\}_1^N$ from the smoothed nonparametric density $\widetilde{f}_{\hat{\epsilon}}(e) = N^{-1}\sum_{i=1}^N K_h(\hat{\epsilon}_i - e)$.
3. Generate boostrapped analogue under $\mathbb{H}_0$, e.g. $Y_i^* = \widehat{a}(X_i^*) + \widehat{h}(V_i^*) + \epsilon_i^*$.
4. Compute the test statistic W^* with the sample $\{(Y_i^*, X_i^*, V_i^*)\}$.

5. Repeat the above steps B times and obtain the bootstrapped test statistics $\{W_j^*\}_1^B$. Compute the bootstrapped p-value $p^* = B^{-1}\sum_j^B \mathbf{1}[W_{j=1}^* > W_N]$ and reject H_0 if p^* is smaller than some prescribed level of significance.

1.6 Finite Sample Results

In this section, I obtain the power and size results using simulations under various data generating processes (DGP). 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 this experiment, I present situations where a more efficient ASF might be available even if the original model is nonseparable in nature.

In each DGP, the number of quantiles is allowed to vary as the choice of P_N 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. Test statistics with and without bias corrections are calculated to compare the usefulness of such techniques in empirical situations. The rule-of-thumb bandwidth of Silverman, i.e. $h = 1.06 \times s.e.(U) \times N^{-r}$, has been implemented. 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.

1.6.1 DGP 1

The data is generated from the following DGP,

$$\mathbb{H}_{11} : Y = (X + \varepsilon) + \delta (X\varepsilon)^2 \tag{1.36}$$

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

$$\begin{aligned} X &= \frac{1}{4} + V - \frac{1}{4}V^2 + u_2 \\ \varepsilon &= \frac{1}{2}V + u_1 \end{aligned}$$

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

Table 1.1 displays the results of empirical size studies under the null $\mathbb{H}_0$, which sets $\delta = 0$ under various number of observations, N, and smoothing options, r_1. Table 1.1 also presents the power analysis under $\mathbb{H}_1$ of DGP 1, as the nonseparability parameter δ varies.

Table 1.1: Empirical Size and Power Results of DGP 1

			$\delta = 0$			$\delta = 0.5$			$\delta = 1$		
N	P_N	BC	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
250	4	N	0.010	0.032	0.058	0.288	0.408	0.476	0.548	0.688	0.736
250	4	Y	0.000	0.006	0.024	0.204	0.368	0.468	0.524	0.652	0.732
250	6	N	0.012	0.036	0.068	0.304	0.412	0.468	0.568	0.700	0.736
250	6	Y	0.000	0.016	0.024	0.212	0.344	0.440	0.508	0.632	0.712
250	8	N	0.004	0.020	0.036	0.320	0.412	0.464	0.580	0.688	0.732
250	8	Y	0.012	0.056	0.092	0.208	0.324	0.396	0.504	0.604	0.664
500	4	N	0.028	0.072	0.108	0.588	0.724	0.776	0.848	0.924	0.940
500	4	Y	0.012	0.040	0.074	0.620	0.772	0.820	0.908	0.948	0.968
500	6	N	0.024	0.068	0.084	0.620	0.736	0.780	0.856	0.924	0.944
500	6	Y	0.012	0.028	0.036	0.624	0.760	0.812	0.908	0.940	0.960
500	8	N	0.028	0.060	0.084	0.640	0.736	0.776	0.860	0.928	0.944
500	8	Y	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 $\kappa_2 = 0.025$. For bias correction (BC), Y=yes, N=no.

Column 4-6 of Table 1.1 display results of the empirical sizes whereas powers are in column 7-12. Table 1.1 displays the results of empirical size and power studies under the

null $\mathbb{H}_{10}$, which is, $\delta = 0$ under various number of observations, N, and number of quantile regions, P. When $\delta = 0$, $\mathbb{H}_0$ is a simple additive model. The first three columns give empirical size results in small and moderate sample sizes, i.e. $N = 250$ or 500. In small samples, our 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 number of Monte Carlo reps is enlarged. 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.[12] 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 $\mathbb{H}_{11}$, tests with analytic variances could deliver reasonable powers, but may depend on the nature of nonseparability in the DGP.

1.6.2 DGP 2

DGP 2 considers more general nonlinearity other than the polynomial forms, specified in (1.37).

$$\mathbb{H}_{21} : Y = X + \varepsilon + \delta \frac{\exp(2X)}{2 + \sin(\varepsilon)} \tag{1.37}$$

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 take various values from $\{0.1, 0.25\}$. When $\delta = 0$, the model goes back to DGP 1 and empirical results are presented in Table 1.1. So only the alternatives need to be studied. When $\delta = 0.1$, the nonseparability is quite weak and it may approximate cases under local power. To this end, I hope to check how the test performs in the adverse cases with relatively small samples.

[12] To save space, more simulation results are not presented in the main text but is available upon request.

Table 1.2: Empirical Power Results under of DGP 2

N	δ	P_N	BC	0.01	0.05	0.1
250	0.1	4	N	0.052	0.120	0.156
250	0.1	4	Y	0.104	0.232	0.372
250	0.1	6	N	0.088	0.140	0.192
250	0.1	6	Y	0.144	0.252	0.344
250	0.1	8	N	0.080	0.128	0.164
250	0.1	8	Y	0.144	0.252	0.336
250	0.25	4	N	0.568	0.748	0.824
250	0.25	4	Y	0.912	0.992	1.000
250	0.25	6	N	0.640	0.772	0.852
250	0.25	6	Y	0.924	0.992	1.000
250	0.25	8	N	0.624	0.744	0.796
250	0.25	8	Y	0.928	0.992	0.996

Note: $N_{mc} = 250$. Smoothing parameters, $r_1 = 1/7.9, r_2 = 1/7.9$. Trimming parameters, $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$. For bias correction (BC), Y=yes, N=no.

The small sample power results are presented in Table 1.2, it is not hard to see that the rejection rates depend heavily on how separable the DGP is. As with $\delta = 0.1$, rejection rates are generally low. Nevertheless, there are still powers even under the almost local alternatives. In contrast, as more weight is put on the nonseparable part, i.e. $\delta = 0.25$, there are quite reasonably large rejection probabilities. The additional nonseparability gives good powers even in such small samples. Finally, I do find that bias correction techniques make a significant impact in leveling up the rejection probabilities.

1.6.3 DGP 3

DGP 3 incorporates multiple unobservables featured from the beginning. For simplicity, assume the true model is like (1.38),

$$\mathbb{H}_{30} : Y = X\eta + \varepsilon \tag{1.38}$$

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 (1.39), denoted by DGP

3.1.

$$\mathbb{H}_{31} : Y = X\eta + \varepsilon + \delta \exp(X\varepsilon) \tag{1.39}$$

Table 1.3: Empirical Size Results of DGP 3

δ	N	P_N	BC	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
0	250	4	N	0.024	0.080	0.100
0	250	4	Y	0.016	0.016	0.032
0	250	6	N	0.032	0.084	0.100
0	250	6	Y	0.016	0.016	0.032
0	250	8	N	0.036	0.076	0.092
0	250	8	Y	0.016	0.016	0.032
0	500	4	N	0.032	0.096	0.132
0	500	4	Y	0.004	0.040	0.076
0	500	6	N	0.048	0.112	0.132
0	500	6	Y	0.016	0.044	0.076
0	500	8	N	0.052	0.108	0.132
0	500	8	Y	0.016	0.040	0.080

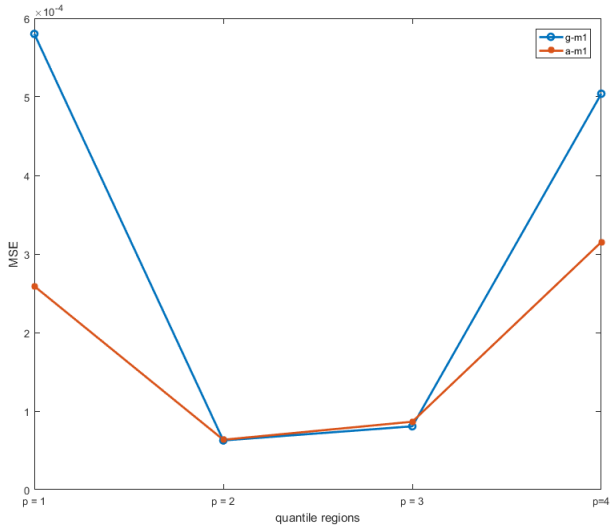
Note: Number of replications ,$N_{mc} = 250$. Smoothing parameters, $r_1 = 1/7.9, r_2 = 1/7.9$. Trimming parameters, $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$. For bias correction (BC), Y=yes, N=no.

The empirical size results are presented in Table 1.3 and power results are in Table 1.4. $\mathbb{H}_{30}$ is one of the example where the condition in Proposition 1.3 holds. It is a nonseparable structural function with more than one unobservable. From Table 1.3, it is true that the test has no power against structural separability. From $\mathbb{H}_{30}$, 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. $\mathbb{H}_{31}$ showcases the situation when more nonseparable forms are added in, 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 unavoidable unless one takes into account the nonseparable nature properly.

However, even though the test has no power against separability under $\mathbb{H}_{30}$, yet it is still useful in that it would yield a more efficient estimator of ASF and its variants. Figure 1.1 plots the mean squared errors (MSE) of ASF estimators with and without

additive restrictions in four quantiles. From the picture, in terms of MSE, the one-step backfitting estimator (solid red) dramatically outperforms the less efficient empirical integration estimator (solid blue), especially in the boundary quantiles.

Figure 1.1: Mean Squared Error of Quantile Average ASF Estimators



Note: Solid red: $\widehat{a}(\cdot)$; solid blue: $\widehat{g}(\cdot)$. 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 $\kappa_2 = 0.025$.

Turn to the empirical power analysis in Table 1.4. The setup of this experiments 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 four dimensions. First, as sample size increases from 250 to 500, the rejection rates increase by about 30%, yielding reasonable powers. Second, the powers do not change much across selected 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. Lastly, I do see that the recursive bias correction techniques make a difference, especially in moderate sized samples.

Table 1.4: Empirical Power Results of DGP 3

			$\delta = 0.5$			$\delta = 1$		
N	P_N	BC	0.01	0.05	0.1	0.01	0.05	0.1
250	4	N	0.440	0.556	0.596	0.628	0.756	0.808
250	4	Y	0.400	0.540	0.616	0.644	0.748	0.792
250	6	N	0.456	0.548	0.584	0.644	0.752	0.788
250	6	Y	0.404	0.504	0.596	0.648	0.720	0.764
250	8	N	0.460	0.540	0.584	0.652	0.748	0.784
250	8	Y	0.384	0.492	0.576	0.624	0.716	0.748
500	4	N	0.756	0.860	0.880	0.856	0.932	0.956
500	4	Y	0.804	0.896	0.928	0.916	0.972	0.976
500	6	N	0.796	0.860	0.884	0.956	0.976	0.980
500	6	Y	0.796	0.892	0.912	0.932	0.964	0.976
500	8	N	0.804	0.864	0.888	0.888	0.940	0.960
500	8	Y	0.800	0.884	0.908	0.956	0.980	0.984

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 $\kappa_2 = 0.025$. For bias correction (BC), Y=yes, N=no.

Although it has been acknowledged by many authors that nonparametric tests with analytical asymptotic variances usually perform poorly in finite samples [115, 50, 86], nonetheless, our limiting variances work reasonably well as shown above. This advantage may translate into great saving of computing time, compared with the commonly used Bootstrapped approaches.

1.7 Extensions

In this section, three extensions are presented to include many commonly encountered empirical cases. Firstly, I address the problematic "curse of dimensionality". Despite all attractive properties that nonparametric testing entails, in real applications, semiparametric index models are often invoked due to the high dimensionality of the control covariates. I accommodate the test in a two-step semiparametric scenario. While the finite dimensional parameters are estimated in the first step by weighted semiparametric least square (WSLS), then the test statistic can immediately be applied on the estimated single index [106, 58, 57, 74, etc.]. I show that the limiting variances under $\mathbb{H}_0$ can be simply obtained by plugging-in the index estimators. In the second extension, when panel data are available, the cross-sectional procedure can be modified to test additivity of time-invariant unobserved

individual-heterogeneity. With repeated observations over time, control variables usually take the form of individual-specific summarized measures, e.g. average value over time, once the exchangability condition holds [7, 94, etc.]. In the last extension, I consider the nonparametric nonseparable triangular simultaneous equations models. Following Imbens and Newey [60], the marginal cumulative distribution function (CDF) of the first stage error suffices to work as a control variable. I show that the asymptotic properties of our test statistic are robust to the problem of "generated regressors".

1.7.1 Semiparametric Test

When the dimension of X (or V) is large in real settings, dimension reduction techniques are often required. 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 parameters. Such models have been studied inPowell et al. [106], Ichimura [57], Powell [105], Ai and Chen [5], Das [29], Klein and Spady [74], Klein and Shen [70], etc.

Now redefine the model (1.1) as the semiparametric single index nonseparable model in Eq. (1.40).

$$Y = m(X'\beta_0, \varepsilon) \tag{1.40}$$

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

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

where the ASFs are defined in Eq. (1.41) and Eq. (1.42), respectively.

$$g(x'\beta_0) = \int m(x'\beta_0, e)dF_\varepsilon(e) = \int E(Y|x'\beta_0, v)dF_V(v) \tag{1.41}$$

$$a(x'\beta_0) \quad = \quad E(Y - h(V)|x'\beta_0), \text{where } h(v) = \int_{\mathbb{R}} E(Y|x'\beta_0, v)dF_{X'\beta_0}(x'\beta_0) - E(Y) \tag{1.42}$$

To conduct the semiparametric inference, one can apply a two-step procedure. In the first step, a consistent estimator of β is obtained by employing the multiple-index WSLS in Ichimura and Lee [58]. 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 1.4.

It is well-known 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.[13] Note that when X and ε are not correlated of any sort, model (1.40) can be rewritten as the semiparametric single index regression with an additive error like Ichimura [57]. 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 $E(U|X) = 0$. As opposed, in the presence of endogenous regressors, by imposing the single index Assumption S-1 in conjunction with Assumption I-1 and I-2, one is still able to work on an additive model such as (1.43).

$$Y = E(Y|X'\beta_0, V) + \epsilon \tag{1.43}$$

where $E(\epsilon|X'\beta_0, V) = 0$ by construction.

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

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

Assumption S.1 only assumes that the equality $E(Y|X, V) = E(Y|X'\beta_0, V)$ holds at the true parameter values. However, it is unlikely to hold elsewhere.

Now consider the consistent estimation of β_0 by WSLS by Ichimura and Lee [58][14]. $\widehat{\beta}$

[13] 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.

[14] Model (1.40) coincides with the generalized regression model in Han [43] if strict monotonicity of m in $X'\beta_0$ is assumed. Han estimate β_0 by maximum rank correlation.

is 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. (1.14). $\widehat{W}_i(\widehat{\beta}^0) = 1/\widehat{E}(\hat{\epsilon}^2|X_i'\widehat{\beta}^0, V_i)$ 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 (1.43) are nested by nonseparable models (1.40), $\widehat{\beta}$ is also consistent even if the true model is additive separable. $\sqrt{N}$-consistency are immediately established in Proposition 1.4 which is a direct implication of the theorem in Ichimura and Lee [58]. See Ichimura [57] and Klein and Shen [70] for details.

Proposition 1.4 (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 simply replace β_0 with $\widehat{\beta}$ and restrict X as a single index, $X'\widehat{\beta}$. Fortunately, the semiparametric covariance estimator, $\widehat{\Omega}_N(\widehat{\beta})$ defined in Eq. (1.44) takes exactly the same form as the nonparametric counterpart. The variability of first-stage estimation of β_0 affects the variance calculation through noting but the single index on X. Note that there is no direct impact. 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 $\sqrt{N}(\widehat{\beta} - \beta_0) = o_p(1)$. Therefore, this test can be considered robust to the first-stage estimator in this aspect. As a consequence, Theorem 1.1, Theorem 1.2 and Corollary 1.2.1 would immediately apply by imposing the single index assumption on the influence function such as Eq. (1.44), a

consistent estimator of which is straightforwardly obtained by plugging in $\widehat{\beta}$ for β_0.

$$\widehat{\xi}_i^p(\beta_0) \equiv [\widehat{t}_i^p + \widehat{E}(t^p|V_i)]\widehat{f}^*(X_i'\beta_0, V_i) - \widehat{t}_i^p]\hat{\epsilon}_i + \widehat{E}(t^p)\widehat{h}(V_i) \tag{1.44}$$

where the quantile (and trimming) function $t_i^p = t^p(X_i'\beta_0)$ is defined over $\mathbb{R}$.

Theorem 1.5 (Asymptotic null distribution)**.** *Under $\mathbb{H}_0$ and Assumption S.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 take advantage of this additional information by solving a nested minimization problem below,

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

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 in Eq. (1.45)

$$\widehat{a}^e(X_i'\widehat{\beta}^a) \equiv \widehat{E}(Y - \widehat{h}^a(V_i)|X_i'\widehat{\beta}^a) \tag{1.45}$$

where $\widehat{h}^a(v)$ is estimated with the more efficient estimator $\widehat{\beta}^a$.

Remark 2. A more powerful test. When the first-stage finite parameters are present, it may open 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, $\beta_a = \beta_0$ only when $m(\cdot, \cdot)$ 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.

1.7.2 Panel Data Test

The second extension applies to situations where panel data is available. Usually panel data consist of same individuals observed over multiple time periods or groups. In the following model, T is assumed finite and let N go to infinity. Suppose the nonseparable panel data model, spanning both cross-sectional and time dimensions, can be specified in Eq. (1.46),

$$Y_t = m_t(X_t, \varepsilon_t),\ t = \{1, 2, \cdots, T\} \tag{1.46}$$

where $m_t : \mathbb{R}^{d_X} \times \mathbb{R}^{\infty} \to \mathbb{R}$ is an unknown time-varying function and the unobservables are very likely to be multi-dimensional, including both time-varying and time-invariant heterogeneity, both of which can arbitrarily interact with X. Altonji and Matzkin [7] study the identification of local average response in this framework via the control function approach and they give conditions on how to generate control covariates, V_t that satisfy Assumption I.1 and I.2 from the panel structure. Essentially, they consider the exchangeability in Assumption P.1.

Assumption-P.1 Exchangeability. $F_{\varepsilon_t|X_1,X_2,\cdots,X_T} = F_{\varepsilon_t|X_{1t_1},X_{2t_2},\cdots,X_{t_T}}$ for $t_i \in \{1, 2 \cdots, T\}$ and $t_i \neq t_j$.

Under this assumption, the error distribution is symmetric in the permutation of X_t. Discussion of the validity of Assumption-P.1 can be found in Altonji and Matzkin [7]. Assumption-P.1' alone cannot guarantee the existence of external control variables. Whereas the implied condition of conditional independence in Assumption-P.1' is the key

to resolving endogeneity.

Assumption-P.1' For each t, $X_t \perp \varepsilon_t | V(X_1, X_t, \cdots, X_T)$, where X_t and ε_t are not measurable with respect to σ-field generated by $V(\cdot)$ which is a known vector of symmetric functions in $(X_1, X_2, \cdots, X_T)$.

Analogous to Maurer et al. [94], V can include individual averages over time, $V = (\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_{d_X})$, where $\overline{X}_j = T^{-1}\sum_{t=1}^{T} X_{jt}$, for each $j \in 1, 2, \cdots, d_X$. In addition, it may include quadratic functions capturing the dispersions, such as $\sum_{t=1}^{T}(X_{jt} - \overline{X}_j)^2$ for each j.

Then back to our testing context, for each time period t, I revise the original set of hypotheses to be $\{\mathbb{H}_0'', \mathbb{H}_1''\}$. By simply replacing V with $V(X_1, X_2, \cdots, X^T)$ in conjunction with Assumption-S.2', all asymptotic results in Section 1.5 trivially hold .

$$\mathbb{H}_0 : g_t(x_t) = a_t(x_t), \text{ a.s., for each } x \in \mathcal{X} \text{and each } t; \ \mathbb{H}_1 : \mathbb{H}_0 \text{ is not true}$$

1.7.3 Test in Triangular Simultaneous Equations Models

In many cases, external control variables V are not observable or available. However, some instrumental variables Z satisfying exclusion restriction may be eligible to provide sources of external variation.

Suppose that the endogenous regressors are determined by first stage equations given in (1.47)

$$X_k = h_k(Z, \eta_k), \quad \forall k = 1, \cdots, d_X \tag{1.47}$$

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. In the example of production function estimation, firm-level cost shifters like input price variation, if observed, can serve as excluded variables as long as they are not correlated with any productivity shocks or factors other than input choices. 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 [60] consider the above assumptions in nonparametric nonseparable triangular simultaneous equations models and prove the following proposition.

Proposition 1.5 (Theorem 1, Imbens and Newey [60]). *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 1.5 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 derived 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. (1.48)

$$\widehat{V}_k = \widehat{F}_{X_k|Z}(x, z) = \frac{\sum_{i=1}^N K_{h_3}(Z_i - z)]\{\mathbf{1}[X_{ki} \leqslant x] - \widehat{\delta}_i(z)\}}{\sum_{i=1}^N K_{h_3}(Z_i - z)}, \; k = 1, 2, \cdots, d_X \tag{1.48}$$

Fortunately, asymptotic results of the test statistic are not influenced by the first stage estimation. Theorem 1.6 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. [87] 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. (1.23) with all V replaced by $\widehat{V}$ and $\widehat{\Omega}(\widehat{V})$ obtained in the similar fashion.

Theorem 1.6 (Asymptotic null distribution)**.** *Under* $\mathbb{H}_0$ *and Assumption T.1-T.2, I.2 and A.1-A.6, then* $\widehat{W}_N(\widehat{V}) \xrightarrow{\text{D}} \widehat{W}_N$.

The proof of Theorem 1.6 exploits the empirical processes arguments. Supporting lemmas can be found in Mammen et al. [87].

1.8 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 excess heterogeneity. As opposed to the previous methods, the test proposed relates the ASF to the additivity of unobservables. The usefulness of ASFs has been suggested by empirical researchers. In this paper, I demonstrate that ASFs contain important information about the additive separability of unobservables and could be extracted for testing purpose. The specification test, in turn, can shed light on the estimation of ASFs in terms of consistency and efficiency. So it can be foreseen that the ASF-based test would have wide applicability.

Besides, not only are the analytic asymptotic variances easy to compute but also it works reasonably well in the finite sample studies. The Monte Carlo results confirm it. To be specific, the EQM test exhibits more power as the underlying model becomes more "nonseparable". 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 left for future research. For empirical applications, I suggest researchers to experiment with several values as a robustness check. As discussed at the end, with slight modifications, the results above extend to other empirical scenarios. Such extensions include semiparametric models, panel data and triangular simultaneous equations.

.1 Proofs of Identification Results

Proof of Proposition 1.1. Given $x \in \mathcal{X}$, it follows that

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

where the conditional expectation function (CEF) is denoted by $C(x,v) \equiv E(Y|X=x,V=v)$ for any pair $(x,v) \in \mathcal{X} \times \mathcal{V}$. The last equality invokes the large support assumption to obtain point identification. The identification result of ASF of nonseparable models is given in Blundell and Powell [15], Imbens and Newey [60], etc. □

Proof of Proposition 1.2. 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 (1.2) 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. [99] for detail.

Integrate marginally with respect to v on both sides,

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

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 1.3. 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 $X = x$,

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

and where

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

□

.2 Immediate Lemmas for Asymptotic Theory

Some notation. 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.

Lemma 1 (U-statistic. Serfling [109])**.** *A "mth-order" U-statistic of the form*

$$U_N \equiv \binom{N}{m}^{-1} \sum_{i_1=1}^{N-m+1} \sum_{i_2>i_1}^{N-m+2} \cdots \sum_{i_m>i_{m-1}}^{N} p_N(W_{i_1}, \ldots, W_{i_m})$$

where p_N is a symmetric in $W_{i_1}, \ldots, W_{i_m}$. Suppose that $\mathbb{E}||p_N(W_i, W_j)||^2 = o(N^{m-1})$. Also define

$$\begin{aligned}
r_N(W_i) &\equiv \mathbb{E}[p_N(W_{i_1}, \ldots, W_{i_m})|W_{i_1}] \\
\theta_N &\equiv \mathbb{E}[r_N(W_i)] = \mathbb{E}[p_N(W_{i_1}, \ldots, W_{i_m})] \\
\widehat{U}_N &\equiv \theta_N + mN^{-1} \sum_{i=1}^{N} [r_N(W_{i_1}) - \theta_N]
\end{aligned}$$

where θ_N is assumed to exist and $\widehat{U}_N$ is called the "projection" of U_N. Then

$$U_N - \widehat{U}_N = o_p(N^{-1/2})$$

Lemma 2 (CEF)**.** *Suppose $E|(Y|U)|^2 < \infty$, $\inf_{\mathcal{U}} f(u) > 0$ where $f(\cdot)$ is the density function of U that is everywhere positive, $d < 4$. For each $u \in \mathcal{U}$, the bias correction estimator $\widehat{C}(u) = \widehat{E}(Y|U = u)$ is defined in Eq. (1.14), then it follows that*

$$a). \qquad |\widehat{C}(u) - C(u)| = f_U(u)^{-1} \left| N^{-1} \sum_{i=1}^{N} K_{i,\cdot}^{U} (\widetilde{Y}_i - C(u)) \right| + o_p(N^{-1/2})$$

$$b). \qquad N^{-1} \sum_{i=1}^{N} K_{i,\cdot}^{U} (\widetilde{Y}_i - C(u)) = N^{-1} \sum_{i=1}^{N} K_{i,\cdot}^{U} \epsilon_i + O(h^4)$$

where $\widetilde{Y}_i = Y_i - \widehat{\delta}_i(u)$, $\widehat{\delta}_i(u) = \widehat{C}_0(U_i) - \widehat{C}_0(u)$ and $\epsilon_i = Y_i - C(U_i)$. $\widehat{C}_0(\cdot)$ is the preliminary conditional expectation estimator in Eq. (1.13).

Proof. For a). one must show that

$$DC = \left| \frac{N^{-1} \sum_{i=1}^{N} K_{i,\cdot}^{U} \widetilde{Y}_i}{\widehat{f}_U(u)} - C(u) \right| \left| \frac{\widehat{f}_U(u)}{f_U(u)} - 1 \right| = o_p(N^{-1/2})$$

Under the well-known nonparametric rate, one would have

$$|\widehat{f}_U(u) - f_U(u)| = O_p(h^2 + N^{-1/2} h^{-d/2}); \left| N^{-1} \sum_{i=1}^{N} K_{i,\cdot}^{U} \widetilde{Y}_i - f_U(u) C(u) \right| = O_p(h^2 + N^{-1/2} h^{-d/2})$$

From this, let $c_f = \sup_u \widehat{f}(u)$,

$$\begin{aligned} \left| \frac{N^{-1} \sum_{i=1}^{N} K_{i,\cdot}^{U} \widetilde{Y}_i}{\widehat{f}_U(u)} - C(u) \right| \quad &\leqslant c_f \left\{ \left| N^{-1} \sum_{i=1}^{N} K_{i,\cdot}^{U} \widetilde{Y}_i - f_U(u) C(u) \right| + C(u) \left| \widehat{f}_U(u) - f_U(u) \right| \right\} \\ &= O_p(h^2 + N^{-1/2} h^{-d/2}) \end{aligned}$$

Then by Cauchy-Schwartz inequality, we have

$$DC \leqslant \sqrt{\left(\frac{N^{-1} \sum_{i=1}^{N} K_{i,\cdot}^{U} \widetilde{Y}_i}{\widehat{f}_U(u)} - C(u) \right)^2 \left(\frac{\widehat{f}_U(u)}{f_U(u)} - 1 \right)^2} = O(h^4 + N^{-1} h^{-d} + N^{-1/2} h^{2-d/2})$$

Suppose $h = O(N^{-r})$, to make sure $\sqrt{N}DC = o_p(N^{-1/2})$ hold, $1/8 < r < 1/2d$ and generally requires $d < 4$.

For b). To begin with, first add and subtract $K^U_{i,\cdot}\delta_i(u)$ and use the fact $\delta_i(u) = C(U_i) - C(u)$,

$$\begin{aligned} N^{-1}\sum_{i=1}^{N} K^U_{i,\cdot}(\widetilde{Y}_i - C(u)) &= N^{-1}\sum_{i=1}^{N} K^U_{i,\cdot}(Y_i - \delta_i(u) - C(u) - \widehat{\delta}_i(u) + \delta_i(u)) \\ &= N^{-1}\sum_{i=1}^{N} K^U_{i,\cdot}(Y_i - C(U_i) - \widehat{\delta}_i(u) + \delta_i(u)) \\ &= N^{-1}\sum_{i=1}^{N} K^U_{i,\cdot}\epsilon_i - N^{-1}\sum_{i=1}^{N} K^U_{i,\cdot}(\widehat{\delta}_i(u) - \delta_i(u)) \end{aligned}$$

And it is true that

$$\sqrt{N}^{-1/2}\sum_{i=1}^{N} K^U_{i,\cdot}(\widehat{\delta}_i(u) - \delta_i(u)) = o_p(N^{-1/2})$$

To see this, remember $\delta_i(u) - \delta_i(u) = O$

$$E[N^{-1}\sum_{i=1}^{N} K^U_{i,\cdot}(\widehat{\delta}_i(u) - \delta_i(u))] = E[K^U_{i,\cdot}(\widehat{\delta}_i(u) - \delta_i(u))] = O(h^2)O(h^2) = O(N^{-4r})$$

$$\text{Var}\left[N^{-1}\sum_{i=1}^{N} K^U_{i,\cdot}(\widehat{\delta}_i(u) - \delta_i(u))\right] = N^{-1}\text{Var}\left[K^U_{i,\cdot}(\widehat{\delta}_i(u) - \delta_i(u))\right] = O(N^{-2}h^{d+1})$$

As long as $1/8 < r < 1/(d+1)$, b). will hold. See Shen and Klein [110] for higher order bias reduction. □

Lemma 3. *Suppose that $R(\cdot)$ is a measurable function defined on $\mathbb{R}^d$ with continuous and bounded second derivatives. t_i is the quantile or trimming indicator defined in (1.21) and the density function $f_X(\cdot)$ satisfies Assumption A-2. For any $x_0 \in \mathcal{X}$, it is true that*

$$E[t_i^p R(X_i)K_h(X_i - x_0)] = t^p(x_0)R(x_0)f(x_0) + O(h^2)$$

where $t^p(x_0) \equiv \mathbf{1}\{x_0 \in \mathcal{X}_0\}$.

Proof. Define the upper bound $q_1 \equiv q_X(\frac{p-1}{P})$ and lower bound $q_2 \equiv q_X(\frac{p}{P})$,

$$\begin{aligned} E[t_i^p R(X_i) K_h(X_i - x_0)] &= \int_{q_2}^{q_1} k(\frac{x - x_0}{h}) R(x) f(x) dx \\ &= \int_{\frac{q_2 - x_0}{h}}^{\frac{q_1 - x_0}{h}} R(x_0 + uh) f(x_0 + uh) k(u) du \\ &= \int_{\frac{q_2 - x_0}{h}}^{\frac{q_1 - x_0}{h}} [R(x_0) f(x_0) + h(R(x_0) f(x_0))' u + h^2 (R(x) f(x))''|_{x^+} u^2/2] k(u) du \\ &= t^p(x_0) R(x_0) f(x_0) + [R(x_0) f(x_0)]' h \int_{\frac{q_2 - x_0}{h}}^{\frac{q_1 - x_0}{h}} u k(u) du + O(h^2) \end{aligned}$$

The second term is $o(h^2)$. □

.3 Asymptotic Proof

Proof of Theorem 1.1. Recall $\widehat{T}_N^p$ in Eq. (1.24) and $\widetilde{T}_N^p$ in Eq. (1.29). 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 .2 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 $\widehat{T}_N^p = \widetilde{T}_N^p + o_p(N^{-1/2})$. Lemma .2 and Lemma .3 are dealing with D_N^a and D_N^g, respectively. Lemma .4 provides the intermediate result for Lemma .5 on D_N^h. 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 1.1 is established with the limiting variance $\Omega_p = E(\xi_i^p \xi_i^{p\prime})$.

□

Proof of Theorem 1.2 and Corollary 1.2.1. According to Theorem 1.1, it is true that $\widehat{T}_N$ follows a P_N-dimensional multivariate normal distribution.

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

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

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,

$$\begin{aligned}|\widehat{D}(X_i) - D(X_i)| &= |[\widehat{g}(X_i) - g(X_i)] + [\widehat{a}(X_i) - a(X_i)]| \leqslant |\widehat{g}(X_i) - g(X_i)| + |\widehat{a}(X_i) - a(X_i)| \\ &= O_p((Nh)^{-1/2})\end{aligned}$$

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,

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

□

Lemma .2 (D_N^a)**.** *Suppose that Assumption A.1-A.6 hold and under* $\mathbb{H}_0$*, then* D_N^a *in Eq.* (1.31) *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 Lemma A 2, 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 Lemma A 3,

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

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

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

□

Lemma .3 (D_N^g)**.** *Suppose that Assumption A.1-A.6 hold and under* $\mathbb{H}_0$*, then* D_N^g *in Eq.* (1.30) *can be written as the following,*

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

Proof. Recall that D_N^g and by definition,

$$\begin{aligned} D_N^g &= N^{-1}\sum_{i=1}^{N} t_i[\widehat{g}(X_i) - g(X_i)] \\ &= \frac{1}{N(N-1)}\sum_{i,j}^{N} t_i\widehat{C}(X_i,V_j) - \frac{1}{N}\sum_{i=1}^{N}\int t_i C(X_i,v)dF(v) \\ &= \frac{1}{N(N-1)}\sum_{i,j}^{N} t_i\left[\widehat{C}(X_i,V_j) - C(X_i,V_j)\right] + \frac{1}{N(N-1)}\sum_{i,j}^{N} t_i\left[C(X_i,V_j) - \int C(X_i,v)dF(v)\right] \\ &\equiv D_N^{g1} + D_N^{g2} \end{aligned}$$

The third equality follows by adding and subtracting $\sum_{i,j}^{N} t_i C(X_i,V_j)/N(N-1)$.

Part a), for D_N^{g1}, by Lemma A 2, it is true that

$$\widehat{C}(X_i,V_j) - C(X_i,V_j) = \frac{1}{N-2}\sum_{k\neq i,j}^{N} K_{k,i}^X K_{K,j}^V [Y_k - C(X_i,V_j)]/f(X_i,V_j) + o_p(N^{-1/2})$$

Substitute this into D_N^{g1} and rewrite it as $D_N^{g1} = \widetilde{D}_N^{g1} + o_p(N^{-1/2})$, so from now on it suffices to work with $\widetilde{D}_N^{g1}$ defined below,

$$\begin{aligned} \widetilde{D}_N^{g1} &= \frac{1}{N(N-1)(N-2)}\sum_{i,j,k}^{N}\frac{t_i}{f(X_i,V_j)}K_{k,i}^X K_{k,j}^V\left[Y_k - C(X_i,V_j)\right] \\ &= \binom{N}{3}^{-1}\sum_{k>j>i}^{N}\sum_{l=1}^{6}\delta_{g1l}/6 \end{aligned}$$

To represent $\widetilde{D}_N^{g1}$ as a third-order U-statistic and where

$$\begin{aligned} \delta_{g11} &= \frac{t_i}{f(X_i,V_j)}K_{k,i}^X K_{k,j}^V\left[Y_k - C(X_i,V_j)\right]; \delta_{g12} = \frac{t_i}{f(X_i,V_k)}K_{j,i}^X K_{j,k}^V\left[Y_j - C(X_i,V_k)\right] \\ \delta_{g13} &= \frac{t_j}{f(X_j,V_i)}K_{k,j}^X K_{k,i}^V\left[Y_k - C(X_j,V_i)\right]; \delta_{g14} = \frac{t_k}{f(X_k,V_i)}K_{j,k}^X K_{j,i}^V\left[Y_j - C(X_k,V_i)\right] \\ \delta_{g15} &= \frac{t_k}{f(X_k,V_j)}K_{i,k}^X K_{i,j}^V\left[Y_i - C(X_k,V_j)\right]; \delta_{g16} = \frac{t_j}{f(X_j,V_k)}K_{i,j}^X K_{i,k}^V\left[Y_i - C(X_j,V_k)\right] \end{aligned}$$

Moreover, by Lemma A 3,

$$E(\delta_{g11}|X_i) = O(h^4); E(\delta_{g12}|X_i) = O(h^4)$$
$$E(\delta_{g13}|V_i) = O(h^4); E(\delta_{g14}|V_i) = O(h^4)$$
$$E(\delta_{g15}|W_i) = t_i f^*(X_i, V_i)[Y_i - C(X_i, V_i)] + O(h^4)$$
$$E(\delta_{g16}|W_i) = t_i f^*(X_i, V_i)[Y_i - C(X_i, V_i)] + O(h^4)$$

Proving the above results is nothing hard but a little tedious. To conserve space, we only show one term and the others follow the same line of reasoning. Now take $E(\delta_{g15}|W_i)$ as an example,

$$\begin{aligned} E(\delta_{g15}|W_i) &= \int \frac{t(x)}{f(x,v)} K(X_i - x)K(V_i - v)\left[Y_i - C(x,v)\right] f(x)f(v)dxdv \\ &= \int_{\{u_1 : t(X_i+u_1h)=1\}} f^*(X_i + u_1h, V_i + u_2h)\left[Y_i - C(X_i + u_1h, V_i + u_2h)\right] k(u_1)k(u_2)du_1du_2 \\ &= t_i f^*(X_i, V_i)[Y_i - C(X_i, V_i)] + O(h^4) \end{aligned}$$

Note that the second equality holds by the transformation of variables, letting $x = X_i + u_1h$ and $v = V_i + u_2h$. The third equality follows from the Taylor expansion on h around $u_1 = u_2 = 0$.

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

$$\begin{aligned} \widetilde{D}_N^{g1} &= N^{-1}\sum_{i=1}^{N} t_i f^*(X_i, V_i)[Y_i - C(X_i, V_i)] + O(h^4) + o_p(N^{-1/2}) \\ &= N^{-1}\sum_{i=1}^{N} t_i f^*(X_i, V_i)\epsilon_i + o_p(N^{-1/2}) \end{aligned}$$

Part b), for D_N^{g2}, one can also rewrite it as a second-order U-statistic,

$$\begin{aligned} D_N^{g2} &= \frac{1}{N(N-1)}\sum_{j,i}^{N} t_i\left[C(X_i, V_j) - \int C(X_i, v)dF(v)\right] \\ &= \binom{N}{2}^{-1}\sum_{j>i}^{N}(\delta_{g21} + \delta_{g22})/2 \end{aligned}$$

where in particular

$$\delta_{g21} = t_i \left[C(X_i, V_j) - \int C(X_i, v) dF(v) \right]; \delta_{g22} = t_j \left[C(X_j, V_i) - \int C(X_j, v) dF(v) \right]$$

It is obvious that $E(\delta_{g21}|W_i) = 0$. But for δ_{g22}, it can be shown that

$$\begin{aligned} E(\delta_{g22}|V_i) &= E_X[t(X)C(X, V_i)] - E[t(X)Y f^*(X, V)] \\ &= E(t)h(V_i) \end{aligned}$$

where E_X is the expectation taken with respect to only X. The second equality is true only under $\mathbb{H}_0$. Also by Assumption A.1-A.3, we have $E|g_{21}|^2 + E|g_{22}|^2 = O(1) = o(N)$, then the standard U-statistic theorem implies that

$$\widetilde{D}_N^{g2} = N^{-1} \sum_{i=1}^{N} E(t)h(V_i) + o_p(N^{-1/2})$$

Finally, combine $\widetilde{D}_N^{g1}$ and $\widetilde{D}_N^{g2}$, then Lemma .3 follows that

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

□

Lemma 4 and Lemma 5 uses U-statistic theorem to analysis $N^{-1}\sum_{i=1}^{N} \widehat{E}(\Delta(V)|X_i)$.

Some 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. $t_i = \mathbf{1}\{X_i \in \mathcal{X}_0\}$. Under $\mathbb{H}_0$, $Y = m_1(X) + h(V) + \epsilon$.

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

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

Proof. Given $V_i = v \in \mathcal{V}$, recall that $\Delta(v) = \widehat{h}(v) - h(v)$ with $\widehat{h}(\cdot)$ in Eq. (34) and $h(\cdot)$ in Eq. (1.11). Following the similar argument in Lemma .3, $\Delta(v)$ can be decomposed into three components,

$$\Delta(v) = \underbrace{N^{-1}\sum_{i=1}^{N}\left[\widehat{C}(X_i,v) - C(X_i,v)\right]}_{\Delta_1(v)} + \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N} C(X_i,v) - \int C(x,v)dF(x)\right]}_{\Delta_2(v)} + \underbrace{E(Y) - \overline{Y}}_{\Delta_3}$$

By Lemma A 2, $\Delta_1(v) = \widetilde{\Delta}_1(v) + o_p(N^{-1/2})$ where $\widetilde{\Delta}_1(v)$ is defined below,

$$\begin{aligned}\widetilde{\Delta}_1(v) &= \frac{1}{N(N-1)}\sum_{i,j}^{N}\frac{K^X_{j,i}K^V_{j,\cdot}}{f(X_i,v)}[Y_j - C(X_i,v)]\\ &= \binom{N}{2}^{-1}\sum_{j>i}^{N}(d_{ij}+d_{ji})/2\end{aligned}$$

where in particular

$$d_{ij} = \frac{K^X_{j,i}K^V_{j,\cdot}}{f(X_i,v)}[Y_j - C(X_i,v)]; d_{ji} = \frac{K^X_{i,j}K^V_{i,\cdot}}{f(X_j,v)}[Y_i - C(X_j,v)]$$

It is straightforward to show that

$$E(d_{ij}|W_i) = O(h^4); E(d_{ji}|W_i) = \frac{f(X_i)K^V_{i,\cdot}}{f(X_i,v)}[Y_i - C(X_i,v)] + O(h^4)$$

The second moment condition $E|d_{ij}|^2 + E|d_{ji}|^2 = O(h^{-4})$ holds trivially according to Assumption A.1-A.3. So by Lemma A 1, we have

$$\Delta_1(v) = \widehat{\Delta}_1(v) + o_p(N^{-1/2})$$

Note that under $\mathbb{H}_0$, $\Delta_2 = N^{-1}\sum_i^N m_1(X_i) - E[m_1(X)]$. So it is true that $\Delta(v) = \widehat{\Delta}_1(v) + \Delta_2(v) + \Delta_3 + o_p(N^{-1/2})$. □

Lemma .5 (D_N^h)**.** *Suppose that Assumption A.1-A.6 hold and under* $\mathbb{H}_0$*, then* D_N^h *in Eq.* (1.32) *can be written as the following,*

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

Proof. To begin with, a result implied from Lemma .4 states that $E(\widehat{\Delta}|X_i) = o_p(N^{-1/2})$. As this can be seen from below,

$$E(\widehat{\Delta}|X_i) = E[\widehat{\Delta}_1(V)|X_i] + E(\Delta_2|X_i) + E(\Delta_3|X_i) + o_p(N^{-1/2})$$

Also, it is not hard to see the following from Lemma .4 that given X_i,

$$\begin{aligned}
E[\widehat{\Delta}_1(V)|X_i] &= E\left[\frac{f(X_j)K_{i,j}^V}{f(X_j,V_i)}(Y_j - C(X_j,V_i))\middle| X_i\right] = O(h^4)\\
E(\Delta_2|X_i) &= N^{-1}[m_1(X_j) - E(m_1(X))] = O_p(N^{-1})\\
E(\Delta_3|X_i) &= -E[\overline{Y} - E(Y)|X_i] = -N^{-1}[E(Y|X_i) - E(Y)] = O_p(N^{-1})
\end{aligned}$$

Therefore, D_N^h can be further decomposed into four components like below,

$$\begin{aligned}
D_N^h = \; & \underbrace{N^{-1}\sum_{i=1}^{N} t_i[\widehat{E}(\widehat{\Delta}_1(V)|X_i) - E(\widehat{\Delta}_1(V)|X_i)]}_{D_N^{h1}} + \underbrace{N^{-1}\sum_{i=1}^{N} t_i[\widehat{E}(\Delta_2(V)|X_i) - E(\Delta_2(V)|X_i)]}_{D_N^{h2}}\\
& + \underbrace{N^{-1}\sum_{i=1}^{N} t_i[\widehat{E}(\Delta_3|X_i) - E(\Delta_3|X_i)]}_{D_N^{h3}} + o_p(N^{-1/2})
\end{aligned}$$

In what follows, only the first three components need to be analyzed separately.

For D_N^{h1}, we can represent it as a third-order U-statistic,

$$\begin{aligned} D_N^{h1} &= \frac{1}{N(N-1)}\sum_{i,j}^{N} t_i f(X_i)^{-1} K_{i,j}^X [\widehat{\Delta}_{1j} - E(\widehat{\Delta}_1|X_i)] + o_p(N^{-1/2}) \\ &= \frac{1}{N(N-1)}\sum_{i,j}^{N} t_i f(X_i)^{-1} K_{i,j}^X \widehat{\Delta}_{1j} + o_p(N^{-1/2}) \\ &= \frac{1}{N(N-1)(N-2)}\sum_{i,j,k}^{N} \frac{t_i f(X_k) K_{i,j}^X K_{k,j}^V}{f(X_i) f(X_k, V_j)} [Y_k - C(X_k, V_j)] + o_p(N^{-1/2}) \\ &= \binom{N}{3}^{-1} \sum_{k>j>i}^{N} \sum_{l=1}^{6} h_{1l}/6 + o_p(N^{-1/2}) \end{aligned}$$

The first equality holds as we can remove the random denominator of $\widehat{E}(\widehat{\Delta}_1|X_i)$ according to Lemma 2. The second equality is because of the fact $E[\widehat{\Delta}_1|X_i] = O(h^4)$. Substitution of $\widehat{\Delta}_1$ gives the third equality. And where in particular,

$$\begin{aligned} h_{11} &= \frac{t_i f(X_k) K_{i,j}^X K_{k,j}^V}{f(X_i) f(X_k, V_j)} [Y_k - C(X_k, V_j)]; h_{12} = \frac{t_i f(X_j) K_{i,k}^X K_{j,k}^V}{f(X_i) f(X_j, V_k)} [Y_j - C(X_j, V_k)] \\ h_{13} &= \frac{t_j f(X_k) K_{j,i}^X K_{k,i}^V}{f(X_j) f(X_k, V_i)} [Y_k - C(X_k, V_i)]; h_{14} = \frac{t_k f(X_j) K_{k,i}^X K_{j,i}^V}{f(X_k) f(X_j, V_i)} [Y_j - C(X_j, V_i)] \\ h_{15} &= \frac{t_k f(X_i) K_{k,j}^X K_{i,j}^V}{f(X_k) f(X_i, V_j)} [Y_i - C(X_i, V_j)]; h_{16} = \frac{t_j f(X_i) K_{j,k}^X K_{i,k}^V}{f(X_j) f(X_i, V_k)} [Y_i - C(X_i, V_k)] \end{aligned}$$

It is easy to see that $E(h_{1l}|W_i) = O(h^4)$ for $l = \{1, 2, 3, 4\}$ and

$$E(h_{15}|W_i) = E(h_{16}|W_i) = E[t(X)|V_i] f^*(X_i, V_i)[Y_i - C(X_i, V_i)] + O(h^4)$$

It is also true that $\sum_{l=1}^{6} E|h_{1l}|^2 = O(h^{-4}) = o(N^2)$ by Assumption A-5. Then standard U-statistic theorem follows,

$$D_N^{h11} = N^{-1} \sum_{i=1}^{N} E[t(X)|V_i] f^*(X_i, V_i) \epsilon_i + o_p(N^{-1/2})$$

For D_N^{h2}, we can also represent it as a third-order U-statistic,

$$\begin{aligned} D_N^{h2} &= N^{-1}\sum_{i=1}^{N} t_i[\widehat{E}(\Delta_2|X_i) - E(\Delta_2|X_i)] \\ &= N^{-1}\sum_{i,j}^{N} t_i f(X_i)^{-1} K_{i,j}^X[\Delta_{2j} - E(\Delta_2|X_i)] + o_p(N^{-1/2}) \\ &= N^{-1}\sum_{i,j}^{N} t_i f(X_i)^{-1} K_{i,j}^X[m_1(X_j) - m_1(X_i)]/N + o_p(N^{-1/2}) \\ &= o_p(N^{-1/2}) \end{aligned}$$

The second equality holds as we can remove the random denominator of $\widehat{E}(\Delta_2|X_i)$ according to Lemma 2. Substitution of Δ_2 gives the third equality.

For D_N^{h13},

$$\begin{aligned} D_N^{h13} &= -N^{-1}\sum_{i=1}^{N} t_i[\widehat{E}(\Delta_3|X_i) - E(\Delta_3|X_i)] \\ &= -N^{-1}\sum_{i,j}^{N} t_i f(X_i)^{-1} K_{i,j}^X[Y_j/N - E(Y|X_i)/N] + o_p(N^{-1/2}) \\ &= o_p(N^{-1/2}) \end{aligned}$$

So, combining all above,

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

□

Proof of Theorem 1.6. The proof of the theorem is via the functional derivative argument almost identical to Mammen et al. [87] who primarily study the local polynomial estimators.

Essentially, based on their arguments, it is possible to show that given $x \in \mathcal{X}$,

$$\begin{aligned} N^{-1}\sum_{i=1}^{N}[\widehat{C}(x,\widehat{V}_i) - C(x,V_i)] &= N^{-1}\sum_{i=1}^{N}[\widehat{C}(x,V_i) - C(x,V_i)] \\ &\quad + N^{-1}\sum_{i=1}^{N}\frac{\partial C(x,V_i)}{\partial v}(\widehat{V}_i - V_i) + o_p(1/\sqrt{Nh^{d_X}}) \\ &= T_1 + T_2 + o_p((Nh^{d_X})^{-1/2}) \end{aligned}$$

Given that C(x,) has bounded partial derivatives at each x, $T_2 = O(N^{-1/2})$ through stochastic expansion and U-statistic projection arguments. T_1 is already studied in Lemma A 2. □

Chapter 2

Identification and Testing of Nonparametric Production Functions without Hicks-neutral Productivity Shocks

2.1 Introduction

Understanding how inputs are related to outputs is a fundamental issue in empirical industrial organization and other fields of economics, see [1]. 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 on productivity, 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, cost efficiency, effects of R&D, to estimate markups, and to evaluate merger impact. As in many empirical applications, reliable estimates of production function parameters are essential to the conclusions that researchers attempt to make. In this paper, I demonstrate that mistakenly assuming Hicks-neutrality of productivity shocks may cause severe biases of structural parameters and productivity estimates. To resolve this problem, I first propose a robust estimator of average output elasticities of fully nonparametric production functions and then exploit the proposed specification test to investigate whether firm-level production functions obey Hicks-neutrality. Empirical evidence shows 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.

The concept of Hicks neutrality was first put forth in 1932 by John Hicks in the book *The Theory of Wages*. A change is considered to be Hicks neutral if the change does not affect the balance of labor and capital in the production functions. In the cross-sectional context, Hicks-neutrality substantially restricts the form of unobserved productivity shocks across

firms, which could be a potential rich and important source of underlying heterogeneity. It further indicates the absence of firm-level unobserved heterogeneity in input substitutability within the industry. Furthermore, it presumes that output-input elasticities are identical for firms that employ the same amount of inputs. Other than the modeling perspective, it poses serious challenges to many commonly employed identification strategies and consistency of production function estimators, which rely critically on the structural separability between input choices and productivity shocks. Last but not least, the wrongly imposed Hicks-neutral technology would cause the distortion of distributions of firm productivity measures, defined as the "Solow" residuals of log production functions. Motivated by the above facts, I suggest an empirically useful test of Hicks-neutrality and show that such a test can be converted into testing additive separability between input choices and multi-dimensional unobservables. In order to control the endogeneity of flexible inputs, such as labor, I extend the proxy variable approach from empirical industrial organization literature, to fully nonparametric nonseparable production functions. Compared with parametric estimation, the nonparametric structural approach adopted here, is not only robust to misspecification but also allows richer firm-level heterogeneity. The estimation and testing results are given for the static model but most of identification results extend to fully dynamic models with additional assumptions on the productivity process.

Recently, there has been a few papers considering non-Hicks-neutral production functions. The method of this paper differs from theirs in that I consider a more general form of production functions, which is fully nonparametric and place no restriction on how productivity enters and are rich in unobserved heterogeneity. The advantage of using such general functional forms at this level ensures that identification and estimation are not driven by any particular parametrization including the multiplicative structure between input amounts and unobservables. Admittedly, the set of identified objects is limited in contrast to previous models with more restrictions. For empirical applications, a specification test should be informative. Therefore, I derive the testable implications of Hicks-neutrality.

I apply the proposed estimation and testing procedure with firm-level panel data of the U.S. manufacturing industry, including all sectors, from 1990 to 2011. The empirical

estimation results of output elasticities (with respect to labor and capital) suggest that controlling for endogenous inputs is crucial. As opposed to previous nonparametric estimates with only exogenous variables, labor elasticity estimates are reduced on average by 12.5% while it increases by only 9.4% for capital, indicating an upward bias on the return-to-scale parameter. On the other hand, mistakenly imposing Hicks-neutral technology could result in an overestimate (underestimate) bias as high as 3.5% for labor elasticity (3.1% for capital elasticity). More importantly, my findings provide an explanation for the well-known phenomenon of constantly decreasing employment in the manufacturing industry. It suggests that there is an obvious trend in the input substitution patterns over time. To be specific, the relative importance of labor has become weaker while capital has been gaining a stronger position in terms of relative output elasticities. This may be largely due to the technological transformation via the mass adoption of computing and electrical equipments during the period of 1993-1998. I also find slightly scaled economies in this period of manufacturing growth. Unsurprisingly, such rapid change of manufacturing technology has also been captured by the proposed test of Hicks-neutrality. Non-technically speaking, the testing results suggest that from the 1990 to 2000, firm-level productivity affects output by changing the "essential technology" (or altering input substitution patterns), rather than simply scaling up output. A possible explanation is that firms are heterogeneous in the speed of adopting new technology. While such effect disappeared after 2002 when most firms have finished the technological transformation. Results do not change much after controlling for sector-specific effects which are not sufficient to mitigate the problem of non-Hicks neutrality.

The contributions of this paper are mainly twofold. Firstly, this paper contributes to the literature of nonparametric identification and estimation of firm-level production functions. In particular, I consider the identification of fully nonparametric production functions beyond Hicks-neutral productivity shocks in both static and dynamic models. The proposed approach differs from Kasahara et al. [64]. They assume that a firm's productivity belongs to a finite number of types. Within each type, the production function is Hicks-neutral. Whereas I rely on the *proxy* variable approach [101, 77, 120, 2] and employ identification strategy for nonseparable models. The estimators of average output

elasticities are robust to non-Hicks neutral productivity shocks. However, the cost is that firm-level productivity cannot be identified at this level of generality. The identification strategy relies on the control function from the literature of nonparametric identification [16, 59, 60, 35, 65, 90, 32, 116, etc.]. This paper also belongs to a growing literature of estimating firm-level production functions. Griliches and Mairesse [37] consider various identification strategies dealing with the endogenous inputs. Blundell and Bond [14] provide a dynamic panel estimator for parametric production functions. Gandhi et al. [36] derive nonparametric identification by exploiting the first order condition to profit maximization. Exploitations of share equation date back to Klein [68], Solow [112]. Huang and Hu [54] and Kim et al. [67] allow for measurement errors in capital. Secondly, to my best knowledge, this is the first paper to question the commonly assumed Hicks-neutral technological shocks and propose an empirical test for it. Nonparametric estimation of production functions without endogenous inputs has been studied in Varian [117], Vinod and Ullah [118], etc.

The rest of the paper is structured as follows. Section 2.2 reviews the notion of Hicks-neutral production functions and discusses the importance of testing such an assumption. Section 2.3 first addresses the non-identification problem of non-Hicks-neutral production functions and then presents identification results of structural parameters in both static and dynamic models. Nonparametric estimators and testing statistics are given in section 2.4. Finally, data and empirical results are presented in section 2.5 and section 2.6, respectively. Section 2.7 concludes this paper.

2.2 Hicks/Non-Hicks-neutral Productivity

Previous studies rely primarily on parametric specifications such as Cobb-Douglas (CD) production functions. Recently the literature has moved onto more flexible functional forms and even nonparametric production functions as micro production data becomes gradually available. This paper takes a further step by considering estimation in the context of fully nonparametric value-added production functions. Standard notations are used from the empirical IO literature and for brevity, omit the cross-sectional subscription $i \in \{1, \cdots, N\}$, denoting each firm (or plant or establishment). Assuming a single-product firm, define the

firm-level fully nonparametric value-added structural production function in Eq. (2.1),

$$Y_t = F_t(L_t, K_t, \omega_t, \varepsilon_t),\ t = 1, 2, \cdots, T \tag{2.1}$$

where Y_t represents some measure of firm-level value-added output of product. K_t denotes the level of capital input at period t and L_t denotes the amount of labor input at period t. The random vector, $(\omega_t, \varepsilon_t)$, contains unobserved factors in which ω_t represents the time-varying heterogeneity in productivity shocks and ε_t denotes idiosyncratic shocks including measurement errors and unexpected *ex post* errors in output, etc. $F_t : \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^\infty \to \mathbb{R}^+$ is the firm-level time-varying unknown production function that maps observed choices and unobserved random factors into the space of final product. Also note that the dimension of ω_t is assumed to be one for convenience but this restriction can be relaxed. In the following analysis, it is required that the time period T be fixed and the number of firms N be large.

In what follows, Eq. (2.1) is also referred to a non-Hicks-neutral production function. The concept of Hicks-neutrality was first put forth in 1932 by John Hicks in his book *The Theory of Wages*. A change is considered to be Hicks neutral if the change does not affect the balance of labor and capital in the firm's production function. This means that the productivity or technological shocks affects output only in a multiplicative way rather than altering input substitutability. To fix idea, now formally define the Hicks-neutrality production functions in Eq (2.2) for each time period t,

$$Y_t = A_t(\omega_t, \varepsilon_t) F_t^1(L_t, K_t),\ t = 1, 2, \cdots, T \tag{2.2}$$

where $F_t^1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ and A_t is an unknown potentially time-varying function of multi-dimensional unobservables, $A_t : \mathbb{R} \times \mathbb{R}^\infty \to \mathbb{R}^+$[1]. In what follows, I refer to $F_t^1(\cdot)$ as the "core technology" for which the functional form is robust to the impact of firm-specific and common shocks. In more aggregate studies, A_t represents the technological changes over time and could be taken as total factor productivity (TFP) or aggregate productivity.

[1]Since Y_t is defined on $\mathbb{R}^+$, it is usually assumed that $A_t(\cdot) > 0$ as $F_t(\cdot) > 0$.

Taking log transformation of Eq. (2.2) for each t,

$$y_t = f_t^1(L_t, K_t) + a_t(\omega_t, \varepsilon_t) \tag{2.3}$$

where lowercase letters, here and in the following, represent the log transformation of original measures, e.g. $y_t = \log Y_t$, $f_t^1(\cdot) = \log F_t^1(\cdot)$. Hicks-neutrality implies that firm unobserved heterogeneity or productivity shocks scale up or down output in a multiplicative way. This further indicates an additive structure of unobservables for the log transformed production function. A coarse measure of firm productivity can be backed out as the "Solow residuals", a_t, by estimating the log production function.

The most famous example of Hicks neutrality is the CD production function below. The CD production function is a special case of a large class of constant elasticity substitution (CES) production functions which also obey the form of Hicks-neutrality [2].

$$Y_t = A_t(\omega_t, \varepsilon_t) L_t^{\beta_L} K_t^{\beta_K} \tag{2.4}$$

where $\beta \equiv (\beta_K, \beta_L)$ are structural parameters measuring output-input elasticities. Estimators of them are usually obtained by estimating the linear regression model by OLS in Eq. (2.5), after log transformation. However, endogeneity arising from input choices would inevitably bias estimates unless corrected measures are taken.

$$y_t = \beta_L l_t + \beta_K k_t + a_t(\omega_t, \varepsilon_t) \tag{2.5}$$

On the other hand, examples of commonly used non-Hicks neutral production functions include labor or capital-augmented production functions (2.6),

$$Y_t = F_t(K_t, A_t L_t) \text{ or } Y_t = F_t(A_t K_t, L_t) \tag{2.6}$$

and random coefficient CD (2.7), where input elasticities are unknown functions of firm heterogeneity, etc. Under this scenario, productivity shocks could impact firms' core

[2] For more general CES production functions, $Y_t = A_t(\beta_L L_t^{\gamma} + \beta_K K_t^{\gamma})^{1/\gamma}$, where $\gamma \leqslant 1$ determines the degree of substitutability and (β_K, β_L) are respective input shares.

technology through one additional channel—altering the relative weights of labor and capital.

$$Y_t = A(\varepsilon_t) K_t^{\beta_K(\omega_t)} L_t^{\beta_L(\omega_t)} \tag{2.7}$$

In the remainder of this section, I present three motivations for relaxing the Hicks-neutral restrictions and proposing a testable hypothesis. In particular, the Hicks-neutrality imposes enormous restrictions on the pattern of firm heterogeneity and this cannot be easily reconciled with intuition and empirical facts. In addition, such restriction poses serious threats to identification strategies based on the additivity of unobservables after log transformation. Finally, I use a trivial simulation example to show that distribution of productivity measures could be severely distorted once the Hicks-neutrality doesn't hold. Recently, there are also empirical evidences questioning the Hicks-neutral technology that may be too restrictive for certain industries.

2.2.1 Restrictions of Firm Heterogeneity

Generally speaking, imposing the multiplicative separability between F_t^1 and A_t, also known as Hicks-neutral technology, can significantly restrict the form of unobserved productivity heterogeneity and input substitution patterns. Here, I attempt to list some modeling restrictions and implications arising from the commonly assumed Hicks-neutrality.

Firstly, Hicks-neutrality of productivity rules out unobserved heterogeneity in input elasticities. By definition,

$$\beta_{L,t} \equiv \frac{\partial Y_t}{\partial L_t}\frac{L_t}{Y_t} = \frac{\partial y_t}{\partial l_t}, \quad \beta_{K,t} \equiv \frac{\partial Y_t}{\partial K_t}\frac{K_t}{Y_t} = \frac{\partial y_t}{\partial k_t}$$

Given the additive structure of unobservables like Eq. (2.3), output-input elasticities can be written as $\beta_{L,t} = L f^1_{L,t}(K, L)$ and $\beta_{K,t} = K f^1_{K,t}(K, L)$, where subscript L and K of $f(\cdot)$ refers to partial derivatives with corresponding arguments. Both are degenerate functions of productivity shocks. It is clear that firm-specific output elasticities are only functions of observed input choices, so it would become constant when conditioning on a given

pair of labor and capital. Due to this restriction, commonly used functional forms such as labor/capital-augmented production functions are also being ruled out. A subsequent implication is that firm-specific return-to-scale measures would be identical as long as they employ the same amount of capital and labor. Because one can approximate the return-to-scale with the sum of capital and labor elasticities, such as Eq. (2.8).

$$RTS_t = \beta_{L,t} + \beta_{K,t} \tag{2.8}$$

In CD production function (2.4), return-to-scale is constant for each firm. The above restrictions are undesirable as the imposition of Hicks-neutrality completely eliminates the unobserved heterogeneity in the modeling of key structural parameters of interest.

Secondly, Hicks-neutral technology significantly restricts input substitution pattens at firm level. To see this, one can examine the rate of technical substitution for each firm, defined the ratio of marginal product of labor and capital, $MP_{L,t}/MP_{K,t}$. Under Hicks-neutral technology, it follows that

$$\frac{MP_{L,t}}{MP_{K,t}} = \frac{\partial Y_t/\partial L_t}{\partial Y_t/\partial K_t} = \frac{F^1_{L,t}(L_t, K_t)}{F^1_{K,t}(L_t, K_t)}.$$

As can be seen, the ratio is free of productivity shocks. It further implies that substitution patterns within production should be identical for firms with same input amounts despite their differences in realized productivity. A similar story is applicable to elasticity of substitution which is another common parameter measuring the degree of flexibility that capital can be substituted for labor. The above property can be very undesirable in industries where firms could substantially differ in input substitutability. This is especially true if productivity shocks alter the "core technology". At least, one should exercise extreme caution when such restrictive modeling condition has to be assumed.

Finally, if one is willing to take some behavioral assumptions, Hicks-neutral shocks could have implications on input expenditures and firm-specific markups given various market structures. In perfect competition, this suggests that profit-maximizing firms have constant input expenditure compositions conditional on input mix (L_t, K_t). This seems a reasonable argument provided that output and input markets are also perfectly competitive. But it

can be hard to validated empirically. To see this, assuming a firm is maximizing static profit taking output price and wage rate as given. Consider the first order condition with respect to its labor choice,

$$Price_t F_{L,t}(K_t, L_t) = Wage_t \Rightarrow \beta_{L,t} = \frac{Wage_t L_t}{Price_t Y_t} = Share_{L,t}$$

Under Hicks-neutral production, $\beta_{L,t}$ is a constant given (L_t, K_t). When firms compete in a homogeneous product market and face common input prices, then output Y_t must be confined to be constant. Thus, it eliminates possibilities of firm-specific idiosyncratic shocks.

In markets of imperfect competition, Hicks-neutrality could also indirectly impose restrictions on markups. In the spirit of Hall [42], De Loecker [30, 31], it is possible to recover firm-level markups provided the existence of a variable input. Consider the first order condition with respect to labor input of firm's static cost-minimization problem,

$$\lambda_t F_{L,t}(K_t, L_t) = Wage_t, \text{ where} \lambda_t \text{ is the Lagrangian multiplier.}$$

where λ_t can be viewed as the marginal cost of production, i.e. $\lambda_t = MC_t$. The markup can be recovered from the ratio of the estimated input elasticities of labor and observed labor expenditure shares.

$$\frac{Price_t}{MC_t} = \frac{\beta_{L,t}}{Share_{L,t}}$$

where the price over marginal cost might shed light on the market power in the process of production. However, when productivity shocks are Hicks-neutral, $\beta_{L,t}$ is free of unobserved productivity heterogeneity. Given the choice of labor and capital, markups is in proportion to the inverse of labor expenditure shares. At the firm level, it amounts to the removal of all unobserved heterogeneity in the markup calculation given input choices and expenditure shares. However, in general, productivity might leverage markups in ways other than simply shifting relative shares of inputs. Therefore, relaxing the Hicks-neutrality is essential, especially in the flexible modeling of unobserved productivity heterogeneity.

2.2.2 Identification with Endogenous Inputs

Estimating firm-level production functions has a persistent interest in empirical industrial organization, trade and related fields. As very early noticed by Marschak and Andrews [89], input choices may be endogenous due to its correlation with productivity shocks because firms usually make input decisions based on the *ex post* productivity shocks [37], and as a result, ordinary least square estimators of log-linear production functions are usually inconsistent.

For example, consider the CD production function after log transformation in Eq. (2.5). Suppose that labor is the only variable input that can be altered according to the value of productivity a_t whereas capital is the fixed input determined in the preceding period. Empirical evidence had shown that without controlling for this correlation, labor coefficients tend to have upward biases as it captures the partial effects of productivity on output. In contrast, capital coefficients are likely to be underestimated, causing almost insignificant coefficients. As a consequence, controlling for endogeneity has become the most important task in many estimation methods.

As driven by this concern, many identification strategies have been proposed to address endogeneity. However, as the following shall point out, many of them only work under the assumption of Hicks-neutrality. To see this, consider the log transformation of some Hicks-neutral value-added production function, Eq. (2.3), i.e. $y_t = f_t^1(L_t, K_t) + a_t$. The log additivity of unobservables serves as the critical foundation for identification. Once a_t can no longer be written as an additive term, inconsistent estimators of structural parameters could be generated due to the nonseparability arising from non-Hicks-neutrality. To illustrate this, the first and most obvious example might be the IV estimator whose consistency depends critically on the additive separability of unobservables. As exploited by [37], using input prices as instrumental variables is only valid under Hicks-neutral errors, even if their variation across firms could be justified as exogenous empirically[3].

Another method is to exploit the dynamic structures by specifying a productivity

[3]The input prices IV estimator has been subjected to criticisms widely because 1). input cost data at micro level are not usually available. 2). the variation might not be adequate, especially when the final product market is homogeneous. 3). it might capture differences in input qualities or even market power in factor markets.

evolutionary process [8, 14, etc.]. Due to the persistence of productivity, the current shock is usually assumed to be a time-varying function of previous shocks and some state variables with an additive orthogonal noise.

$$\omega_t = h_t(\omega_{t-1}, K_t) + \xi_t,\ t = 1, 2, \cdots, T \tag{2.9}$$

where $\omega_0 = 0$ for normalization. The identification strategy of the dynamic panel method relies on the orthogonality between idiosyncratic shocks $(\varepsilon_t, \varepsilon_{t-1}, \xi_t)$ and past differences in input choices. In addition to the Hicks-neutral technology, it requires the productivity process to be AR(1), i.e. $\omega_t = \rho\omega_{t-1} + \xi_t$. In the case of CD production functions, the first difference, i.e. $y_t - \rho y_{t-1}$, implies a dynamic panel regression free of productivity shocks.

$$y_t = \rho y_{t-1} + \beta_L(l_t - \rho l_{t-1}) + \beta_K(k_t - k_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1}) + \xi_t$$

Lagged and differenced values of static and dynamic inputs are qualified as IV satisfying the orthogonality condition. However, the elimination of productivity terms is only valid when it is additive. Besides, this approach also requires strong parametric assumption and it is not trivial to extend to models beyond linear structures. Furthermore, due to the insufficient variation of changes in capital stocks, capital coefficient estimates, as a results, tend to be lower and hardly significant.

Recently, Gandhi et al. [36] have exploited the revenue share equation of profit-maximizing firms in perfect competition[4]. Given the Hicks-neutral technological shocks, they show that production functions in (2.2) is nonparametrically identified. Now suppose that $A_t = \omega_t$ for simplicity and consider the first order condition (FOC) in log form with the Hicks-neutral shocks from Eq. (2.2).

$$\ln Wage_t = \ln Price_t + \ln F^1_{L,t}(L_t, K_t) + \omega_t$$

where $F^1_{L,t}$ denotes the partial derivative of F^1_t with respect to labor. It follows that ω_t

[4]This approach dates back to Klein [68] and Solow [112] who consider the first order condition with respect to variables inputs.

becomes additive. Then by subtracting the above log input revenue share equation from the log production function, the productivity term will vanish.[5] Note that this observation would not hold in general for non-Hicks-neutral production functions.

Subsequent work by Kasahara et al. [64] also exploits the input revenue share equation but attempts to relax the Hicks-neutral restriction. In particular, they assume a firm's production function belonging to one of J-finite types characterized by the productivity process, but within each type, still complies with Hicks-neutrality.

Finally, a vast amount of empirical work relies on the proxy variable approach [101, 77, 120, 2, etc.]. Essentially, it requires a proxy function, possibly unknown, for productivity. Such functions can be obtained $\phi_t(k_t, m_t)$, where m_t is the log of control variable such as investment or intermediate input. In addition, assume the productivity dynamics is first-order Markovian, i.e. $\omega_t = h(\omega_{t-1}) + \xi_t$. One can substitute this proxy back into the log CD function, i.e.$y_t = \beta_L l_t + \Phi_t(k_t, m_t) + \varepsilon_t$. Then nonparametrically regress $\omega_t = \Phi_t(k_t, m_t) - \beta_K k_t$ on itself from last period and explore the following orthogonality conditions.

$$E\left[\left.\begin{pmatrix}\varepsilon_t\\ \xi_t\end{pmatrix}\right| \mathcal{I}_t; \boldsymbol{\beta}\right] = 0, \text{ where } \mathcal{I}_t = \{K_t, (L, K, M)_0^{t-1}\}.$$

Either two-step or jointly GMM approach with the above moment conditions can used for estimation of parameters pertaining to economic primitives. Most of previous developed estimators are in the context of CD production functions. In the section 2.3, I extend identification results of the proxy function approach to fully nonparametric production functions beyond Hicks-neutrality at the cost of a slightly stronger set of assumptions. The identification also suggests a convenient testing procedure for Hicks-neutral technological shocks, which will be implemented in the empirical section 2.6.

2.2.3 Measures of Productivity

One of the most important aspects of estimating production functions is to extract measures of productivity. By studying the re-distribution of firm-specific productivity, one can

[5] Gandhi et al. [36] discuss detailed steps in recovering nonparametric production functions.

examine the reallocation effects of particular trade or industrial policies, and identify the sources of productivity growth. Productivity is also interpreted as profitability or innovation potentials in many studies that focus on its relation with R&D expenditure. Therefore, a robust and consistent measure of productivity is a crucial premise upon which all the follow-up studies rely. Previously, a common practice is to estimate or approximate the firm-specific productivity as the exponential of the "Solow" residuals from the log regression, in Eq. (2.3). In the case of CD production functions, as Eq. (2.5), the productivity is approximated as $\exp(\widehat{a}_t)$, where $\widehat{a}_t$ is a consistent estimator of the "Solow" residual. Admittedly, this is a coarse measure of productivity unless one can separate the true heterogeneity, $\exp(\omega)$ from the *ex post* shocks, $\exp(\varepsilon)$[6]. However, the presence of the non-Hicks-neutral productivity shocks would not only pose difficulty in the interpretation of using a summarized productivity measure, but significantly undermine its validity, as shown in the example below.

Example. Consider the random coefficient CD production functions in log, a non-Hicks-neutral one. For simplicity, I focus on the cross-sectional production functions without endogenous inputs, i.e. both labor and capital are static and fixed inputs. For the purpose of demonstration, I also assume away the idiosyncratic shocks so that the only unobservable is the firm-specific productivity shock. Remember that adding more unobservables or shocks does nothing but make the interpretation hard, when only exogenous inputs exist.

$$y_t = \beta_L(\omega_t) l_t + \beta_K(\omega_t) k_t$$

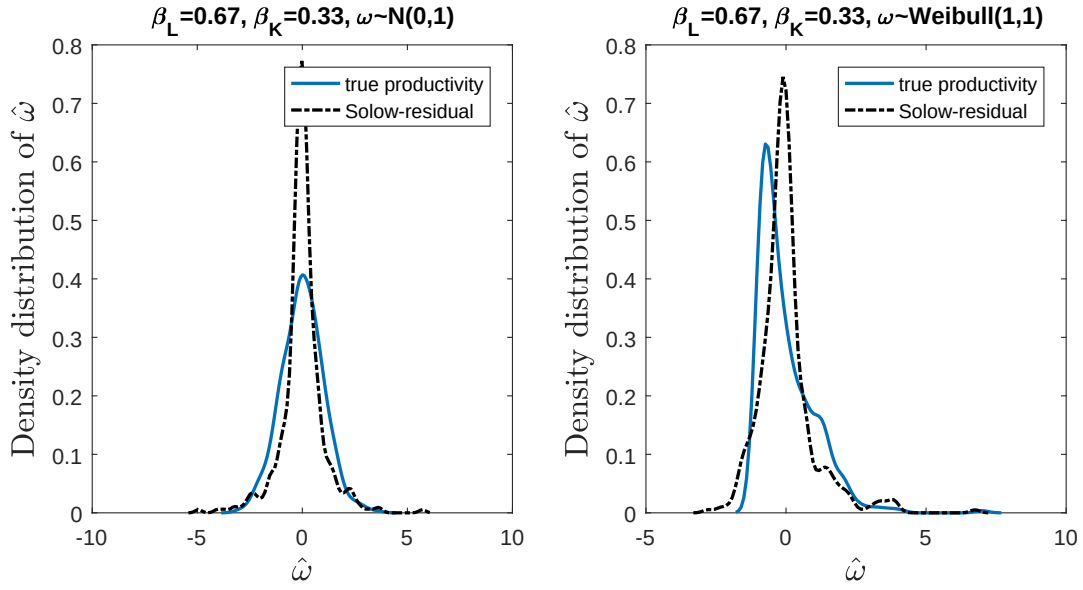
Clearly, it is trivial to note that the distribution of productivity ω_t, if not degenerate, is different from the residuals that are uniformly 0 given any consistent estimator of random coefficients for each firm. A more interesting question is on how the estimated residual distribution looks like compared to the true distribution if estimated using OLS. Intuitively, it depends on the functional form of β_L and β_K as well as the distribution of firm heterogeneity and inputs. To get a first glimpse, I demonstrate it through the following

[6] Olley and Pakes [101] discuss the benefits and costs of using a course measure $\exp(a_t)$ rather than the precise one, $\exp(\omega_t)$.

simulation studies.

Suppose that a simple random coefficient CD production function with $\beta_L = 0.67 + \omega$, $\beta_K = 0.33 + \omega$, where ω is generated from either $N(0,1)$ or Weibull$(1,1)$. I use actual observations for labor and capital from the empirical application. The detailed discussion of data will be postpone in Section 2.6. For simplicity, attention is restricted to the sector of computer and electrics in 2011. The log output is generated according to $y = \beta_L l + \beta_K k + \varepsilon$, where $\varepsilon \sim N(0,1)$. In Figure 2.1, I plot the nonparametric kernel density estimates of true productivity versus estimated Solow-residuals. It is not hard to see that when the true productivity is normally distributed, Solow-residuals tends to display much higher than normal kurtosis. Turning to the right panel of Figure 2.1, the estimated Solow-residuals might have distorted skewness when the true is Weilbull distributed. This example is used to illuminate one fact: severe distortions might occur to the distribution of residual productivity estimates in the presence of non-Hicks-neutral technological shocks.

Figure 2.1: Simulated Productivity Distributions



Note: 1. Total number of observations is 459. 2. Standardized values are plotted to ease comparison.

2.3 Identification of Identifiable Structural Parameters

In this section, I consider the identification issues without Hicks-neutral technology. To begin with, notice that the non-Hicks-neutral production function can be seen as a special case of nonparametric nonseparable models. Thus, borrowing the results from the nonparametric identification literature, it is straightforward to see that structural functions, here production functions, are not identified in the presence of unobservables with unknown dimensions. This further indicates that Hicks-neutrality is generally non-testable without restrictions on unobserved productivity heterogeneity. Alternatively, average production functions are identified. In the second and third subsections, formal assumptions are laid out in both static and dynamic models. I extend the idea of the proxy variable approach in order to identify firm-level average structural parameters such as output-input elasticities and return-to-scales.

Now I cast the above problem into a general econometric question. To begin with, notations are simplified so that the identification problem can be highlighted. Analogous to the log transformation of Eq. (2.1), let the single equation nonseparable model be in Eq. (2.10) where arbitrary interactions between observed and unobserved covariates, e.g. X versus ε, are permitted.

$$Y = m(X, \varepsilon) \tag{2.10}$$

where the unknown measurable function $m : \mathcal{X} \times \mathcal{E} \to \mathbb{R}$ is called the structural function representing primitive economic relations. Nonparametric nonseparable models have been gaining popularity in theoretical econometric works for the past decades. Such models are capable of capturing both the observed and unobserved heterogeneity in structural parameters of important economic interest. For the application here, model (2.10) could represent the log of a nonparametric production function, where Y denotes the log 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 unobservables pertaining to demand and cost conditions.

In contrast, the competing class of specifications, a subset of Eq. (2.10), features the additive separable structure in which the unobservables can be collectively written as an added term, like Eq. (2.11),

$$Y = m_1(X) + m_2(\varepsilon) \tag{2.11}$$

where $m_1(\cdot)$ is a measurable unknown function of only observables defined on $\mathcal{X}$. The above model corresponds to the Hicks-neutral production functions after log transformation, such as Eq. (2.3). In the literature nonparametric specification test of additive separability, the following hypotheses have received a lot of attention.

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

Testing additive separability is a long-standing interest to empirical researchers, especially in structural econometrics. The motivations for the testing hypotheses $\mathbb{H}_0^*$ against $\mathbb{H}_1^*$ are mainly fourfold. 1). This is a test on the absence of unobserved individual heterogeneity in structural functions. Once $\mathbb{H}_0$ is rejected, it implies the partial effect of X on Y is deterministic given the level of observed covariates. As mentioned in Section 2.2, the Hicks-neutral technology, implied by additive separability, imposes a very strong restrictions on the input substitution patterns. 2). Consistency and efficiency of many estimators depend crucially on the separability of disturbances, such as IV estimators. Hahn and Ridder [41] show that the conditional mean restriction, often assumed in IV methods, only has identification power when the model is additive in unobservables. 3). Additive separability of unobservables is often employed in structural models to facilitate identification and estimation. For example, estimating demand using discrete choice models with market-level data, idiosyncratic tastes are usually assumed to be type-I extreme distributed and additive in the mean utility function in order to obtain the closed-form market share equations. Another similar example concerns the unobserved firm fixed cost to be additive in order to

generate the conditional choice probabilities in the dynamic games with entry and exit.[7]

A notable fact is that heterogeneity in microeconomics is rarely unit-dimensional and quite often even the number of dimensions is not known *a priori*. As can be seen, ε often represents unobserved heterogeneity of consumer tastes, product attributes, productivity shocks, measurement errors, etc. Browning and Carro [19] even argue that most empirical models allow less than the maximal amount of unobserved heterogeneity.

As a consequence, testing the additive separability turns out to be challenging, even impossible when no restrictions are placed on the unobservables. Gu [39] studies the identification problem associated with multi-dimensional unobservables (or excess unobserved heterogeneity) in nonseparable structural models. More importantly, he argues that the original hypothesis is not testable to the extent that unobserved heterogeneity is allowed to be modeled as flexibly as possible.

In the following, I first briefly review the non-identification problem of non-Hicks-neutral production functions and give assumptions and results on the identifiable objects—average structural production functions.

2.3.1 Non-identification of Production Functions

First, let us take a look at a simple example mentioned in Benkard and Berry [12]. 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

$$Y = \frac{X}{\sqrt{X^2+1}}\varepsilon_1 + \frac{1}{\sqrt{X^2+1}}\varepsilon_2$$

and

$$Y = \varepsilon_3$$

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

[7] In some structural models, testing separability could yield implications on testing endogeneity, a point taken from Imbens [59] and Imbens and Newey [60].

straightforward to show $Y \sim N(0,1)$ and is independent of X in both specifications. Formally, we follow the definition in Roehrig [107] and let $F_{X,\varepsilon}$ be the distribution and in conjunction with the structure S be a pair $(F_{X,\varepsilon}, \theta)$ that define the data generating process, where θ could denote finite and infinite parameters.

Definition 2.1. *Let F and F' be the distribution functions of (X, Y) implied by the structure S and S', then S and S' are observationally equivalent if $F = F'$.*

Definition 2.2. *The structure S is identified if there is no other S' that is observationally equivalent to S.*

The example above indicates that the structural function itself is not identified with no restrictions on functional forms or distributions, even given all the observed information on (Y, X). More importantly, it also indicates that our original hypothesis of $\mathbb{H}_0^*$ versus $\mathbb{H}_1^*$ is not testable in general because both nonseparable and separable models can give out the same underlying data generating process, meaning they are observationally equivalent.

One solution is to impose additional restrictions to achieve identification [see 91]. 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. [115] assumes that the error term is unidimensional and $m(x, \varepsilon)$ is strictly monotonic for each x. Under such conditions, they arrive a consistent test which also has power against strictly monotonicity if it doesn't hold by taking derivative of the identified structural function. Lu and White [86] transform the original hypothesis into a conditional independence condition. However, they lose equivalence unless certain polynomial functional forms or scalar monotonicity in unobservables is assumed.

However, in the context of this paper and others, it would be very undesirable to assume the unobservables is of single dimension. As already pointed, there are so many unobserved factors that could impact the output quantity beyond the persistent productivity shocks. Even if one assumes that the productivity shocks can be characterized by a scalar term, other idiosyncratic disturbances may not. Normally speaking, even the dimension of ε can be hardly known, let alone the monotonicity shape restriction.

Another direction is to consider the identifiable structural objects instead of seeking

the structural function itself, in the presence of excess unobserved heterogeneity as in Gu [39]. Since the identified parameters themselves are sufficient to answer questions of interest, therefore it would be unnecessary to undertake more assumptions so as to recover the economic primitives. There are many such identifiable parameters available in the nonparametric identification literature. InBlundell and Powell [16], Imbens and Newey [60], they study the *average structural function* (ASF) for which the structural function is integrated with respect to the marginal distribution of unobservables, potentially of unknown dimension. Provided the availability and sufficient variation of control variables, they identify the ASF by integrating out marginally the control covariates of the conditional expectation functions of outcomes. Altonji and Matzkin [7] provide identification and estimation results of an analogous object, *local average response*—the average derivative of X on Y weighted by the conditional distribution of unobservables given X. Hoderlein and Mammen [46] show that the average marginal effects conditional on the observables and quantiles of response, termed *local average structural response*, can be identified.

In this paper, I mainly focus on the identification of ASF and its derivatives as those measures are closest to the single summarized parameters that have been estimated in previous works. But other measures can be obtained, likewise, depending on the research questions. For production functions, define the following structural parameters of interest. Let $F_{\omega_t,\varepsilon_t}$ be the cumulative distributional functions.

Definition 2.3. *Average structural log production functions (ASLPF)*

$$\bar{f}_t(K,L) \equiv \int f_t(K,L,w,e)dF_{\omega_t,\varepsilon_t}(w,e) \tag{2.12}$$

Definition 2.4. *Average structural output-input elasticities (ASOE) of labor and capital*

$$\beta_{L,t}(K,L) \equiv \int \frac{\partial F_t(K,L,w,e)}{F_t(K,L,w,e)}\frac{L}{\partial L}dF_{\omega_t,\varepsilon_t}(w,e) \tag{2.13}$$

$$\beta_{K,t}(K,L) \equiv \int \frac{\partial F_t(K,L,w,e)}{F_t(K,L,w,e)}\frac{K}{\partial K}dF_{\omega_t,\varepsilon_t}(w,e) \tag{2.14}$$

Definition 2.5. *Average output-input elasticities of labor and capital*

$$\bar{\beta}_{L,t} \equiv E[\beta_{L,t}(K,L)] \tag{2.15}$$

$$\bar{\beta}_{K,t} \equiv E[\beta_{K,t}(K,L)] \tag{2.16}$$

In particular, $\bar{F}_t$ is the mean production function, averaging over the cross-sectional distribution of all unobservables and is essentially the ASF considered in Blundell and Powell [16], Imbens and Newey [60]. $\beta_{L,t}$ and $\beta_{K,t}$ represent the average output-input elasticities, with respect to labor and capital across firms. Average output-input elasticities depend on input choices and are potentially heterogeneous across firms. At the same time, they are summarized measures as all correlated and uncorrelated unobservables are marginally integrated out over firms. Counterfactuals would be easily generated once the estimators of the functions are available. Also note that the average return-to-scale (RTS) parameters can be backed out by the addition of $\beta_{K,t}$ and $\beta_{L,t}$.

Note that all above measures have the following properties. First, they are structural objects. It means that any change due to the change of observables like labor and capital choices can be taken as useful counterfactuals as the distribution of unobservables is being held fixed. Second, those objects capture the observable heterogeneity across firms and time periods. Two sectors may give rise to the same average but may differ a lot in the choices of inputs. ASOEs will be able to capture such disparity. Third, given identification, those objects would not be driven by functional forms of production functions, even robust to the Hicks-neutral assumptions. This can significantly reduce the requirements for parametrization and provide means for specification tests against commonly used parametric functions, such as CD production functions.

2.3.2 Identification of Static Models

This subsection considers nonparametric identification of production functions beyond Hick-neutral productivity shocks for the static model. In static models, one does not have to consider dynamic issues such as selection and entry-exit, etc., and therefore it is easy to highlight the identification strategy. In the next subsection, I extend the similar approach

to fully dynamic models together with additional assumptions on productivity evolutionary process. Since with unrestricted multi-dimensional unobservables, structural production functions cannot be point identified, instead, I seek to identify the average objects defined previously as ASLPF and ASOE for each period. The identification can be seen as a nonparametric generalization of the proxy variable Olley and Pakes [101], Levinsohn and Petrin [77], Wooldridge [120], Ackerberg et al. [2].

Now consider the simple static value-added production function with only cross-sectional dimension like in Eq. (2.1). Note that even though the identification and estimation are static, the production function should satisfy a dynamic programming problem with the Bellman equation (2.17).

$$V_t(\omega_t, K_t) = \max\left\{\phi, \sup_{l_t, i_t > 0} \pi_t(\omega_t, l_t, K_t, i_t) - C_t(i_t) + \beta E[V_{t+1}(\omega_{t+1}, K_{t+1})|\mathcal{I}_t]\right\} \quad (2.17)$$

where ϕ is the scrap value upon exit; β is the discount factor; $\mathcal{I}_t$ is the information set at period t; $C_t(\cdot)$ is the investment cost function. The functional form of the production function is allowed to vary each period to capture the time effect. In the static production function, capital K_t is considered predetermined and fixed at period t, whereas labor L_t is the only endogenous input choice. We distinguish two types of unobservables. The scalar-valued ω measures unobserved heterogeneity in productivity and is known to each firm before input choices are made. As opposed, ε represents idiosyncratic or *ex post* shocks, which are orthogonal to input choices and are very likely to be multi-dimensional.

Inspired by the literature on proxy variables, we augment our information set and assume the availability of the proxy such as the investment amount or intermediate inputs (like electricity usage, materials amount, etc.). Similar to Levinsohn and Petrin [77], consider the intermediate demand function, $M_t = \mathbb{M}_t(K_t, \omega_t)$ where $\mathbb{M}_t(k, \cdot)$ is strictly monotonic for each k and each period. We assume that firms are facing common input prices which are time-varying and captured by the functional form of $\mathbb{M}_t$. With the above additional restrictions, Levinsohn and Petrin [77] show that the proxy variable for productivity is obtained by inverting the intermediate demand function, $\omega_t = \mathbb{M}_t^{-1}(K_t, M_t)$ and they substitute this

unknown proxy variable back into the log-transformed CD production function in (2.5).

$$y_t = \beta_L l_t + \beta_K k_t + \mathbb{M}_t^{-1}(K_t, M_t) + \varepsilon_t \tag{2.18}$$

where l_t and k_t are logged labor and capital inputs, respectively. Using arguments specific to partially linear models in conjunction with the productivity process in panel data settings, they show that input shares (β_L, β_k) and the productivity ω_t are identified. Ackerberg et al. [2] address the functional dependence problem in this approach and provide reasonable timing assumptions when labor L_t still has independent variations even after K_t and M_t are conditioned on. Ackerberg et al. [2] consider the conditional intermediate input demand function, i.e. $M = \mathbb{M}(L, K, \omega)$. In general, as long as the production function is Hicks-neutral like in (2.2), by log transformation, one is able to have an additive separable model in the productivity and idiosyncratic errors under which the proxy can be substituted for ω_t such as (2.19).

$$y_t = f_t(K_t, L_t) + \mathbb{M}_t^{-1}(K_t, M_t) + \varepsilon_t \tag{2.19}$$

Once Hicks-neutrality is relaxed, generally we have a nonseparable production function even after the log transformation like (2.20). So proxy variable methods would fail to work when the log production function is nonseparable in unobservables.

$$y_t = f_t(K_t, L_t, \omega_t, \varepsilon_t) \tag{2.20}$$

Proposition 2.1 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. Employing an analogous conditional independence argument with the "generated" control, useful objects such as average production functions, average input elasticities, etc., (defined later) are hence identified. Before moving to identification results, assumptions A-S.1 to A-S.3 formally states the functional form, timing and identification assumptions needed,

A-S.1 Production functions. $Y_t = F_t(K_t, L_t, \omega_t, \varepsilon_t)$, where $\omega_t \in \mathbb{R}, \varepsilon_t \in \mathbb{R}^{\infty}$.

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

A-S.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 $k \in \mathcal{K}$

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

A-S.1 is the functional form assumption. A very general form of production functions is considered. Almost no shape restrictions or distributional assumptions are imposed except for the scalar value of ω_t. However, even this restriction can be relaxed if multiple intermediate inputs are observed. Relaxing this dimensional assumption is beyond the scope of this paper. A-S.2 reiterates the static nature of the current model and highlights the source of endogeneity which is through the interaction between labor choice and productivity. It also simplifies the dynamic process to fit in the cross-sectional studies here. This amounts to treating capital as predetermined and unrelated with the contemporaneous productivity shock. For example, capital could be accumulated deterministically according to $K_t = k(I_{t-1}, K_{t-1})$. A-S.3 is standard in the proxy variable literature. 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. Relaxing this assumption is beyond the scope of the current paper and interesting readers are referred to Huang and Hu [54], Kim et al. [67] who consider cases when capital is measured with errors. Due to the static nature, production functions can be estimated period by period without specifying the productivity evolution process and other dynamic features.

Proposition 2.1 (Conditional Independence)**.** *Under Assumption A-S.1-A-S.3, 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*.*

The proof of Proposition 2.1 resembles Proposition 1 in Imbens and Newey [60] who consider general nonparametric triangular simultaneous equations. But they do differ slightly. In Imbens and Newey [60], the axillary equation is a reduced form of the endogenous

variable on all exogenous and excluded variables. Whereas Proposition 2.1 generates a valid control variate by introducing another endogenous variable which is excluded from the structural function in the current context. The control variable can be estimated by any nonparametric estimators in principle. In Section 2.4 and the empirical application, I consider a bias-reduced local constant estimator.

As for the identification of ASLPF, ASOE and as such, a support condition in A-S.4 is required to generate sufficient variations after conditioning on the control covariate V_t. A-S.4 resolves the functional dependence problem discussed in Ackerberg et al. [2] who provide underlying modeling assumptions to justify the additional variation of labor after conditioning on capital and intermediate inputs. Without the large support, objects such as ASOEs may be only partially identified. In the following, I take A-PF.4 as given but this assumption need to be checked on a case-by-case basis in practice. In Theorem 2.1, ASLPF in Eq. (2.12) and ASOE (2.13) are identified by averaging the conditional expectations weighted by the marginal density of the control covariate by construction.

Theorem 2.1 (Identification)**.** *Under Assumption A-S.1 to A-S.4,* $\bar{f}_t(\cdot)$ *(ASLPF),* $\beta_{L,t}(\cdot)$ *and* $\beta_{K,t}(\cdot)$ *(ASOE) are identified at each* $(K, L) \in \mathcal{K} \times \mathcal{L}$ *for each time period* $t \in \{1, 2, \cdots, T\}$*,*

$$\bar{f}_t(K, L) = \int E_t(y|K, L, v) dF_{V_t}(v) \tag{2.21}$$

$$\beta_{L,t}(K, L) = L \int \frac{\partial E_t(y|K, L, v)}{\partial L} dF_{V_t}(v) \tag{2.22}$$

$$\beta_{K,t}(K, L) = K \int \frac{\partial E_t(y|K, L, v)}{\partial K} dF_{V_t}(v) \tag{2.23}$$

Note that the per-period conditional expectation function along with its first order derivatives can be estimated from data by many nonparametric methods, such as kernel, local polynomial, sieve estimators, etc. So is $F_{V_t}(\cdot)$, the distribution of the control V_t in Proposition 2.1. Also note that $F_{V_t} \sim \text{Uniform}[0, 1]$, so one can also simply $dF_{V_t}(v)$ as dv/v. The identification then is clearly established by construction. The detailed proof of Theorem 2.1 is given in the appendix.

2.3.3 Identification of Dynamic Models

Now it is time to extend the above identification to fully dynamic models in accordance with the mainstream literature. However, identification results only hinge on cross-sectional insights rather than dynamic structures. As with prior studies, we incorporate the definition of the fully nonparametric production function, timing of input decisions, and evolution of productivity processes in the list of identification assumptions stated below.

A-D.1 Functional form. $Y_t = F_t(K_t, L_t, U_t)$, where $U_t = (\omega_t, \varepsilon_t) \in R \times R^{\infty}$ denote multi-dimensional unobservables and ω_t denotes the time-varying unoberved productivity shock; $(K_t, L_t) \in \mathcal{K}_t \times \mathcal{L}_t$.

A-D.2 Timing. K_t is determined in $t-1$ and evolves according to $K_t = \mathbb{K}_t(K_{t-1}, I_t)$; (L_t, M_t) are determined at t.

A-D.3 Scalar monotonicity. The intermediate input function, $M_t = \mathbb{M}_t(k, \omega_t)$ is strictly monotonic in ω_t for each $k \in \mathcal{K}$ and each t.

A-D.4 Markov Productivity. i). $P(\omega_t|\mathcal{I}_{t-1}) = P(\omega_t|\omega_{t-1})$, where the information set at $t-1$ is $\mathcal{I}_{t-1} = (K_t, \{Z_{t-1}\}_2^t)'$ and $Z_t = (K_t, L_t, M_t)'$; ii). $\omega_t = \mathbb{W}_t(w, \eta_t)$ is strictly monotonic in η_t for all w. iii). $F_{\varepsilon}(\varepsilon_t|\mathcal{I}_{t-1}) = F_{\varepsilon}(\varepsilon_t)$.

A-D.5 Initial condition. $\omega_0 = \mathbb{W}_0(K_0)$, where $\mathbb{W}_0 : \mathcal{K} \to \mathcal{W}$ is a strictly increasing measurable function.

Assumptions A-D.1 to A-D.4 resembles those used in static models with very slight modifications. Basically, A-D.1 specifies the fully nonparametric value-added dynamic production functions of interest. The only restriction, which can relaxed, requires a single-dimensional time-varying persistent unobserved heterogeneity. But the idiosyncratic shocks are assumed to be fully flexible. A-D.2 gives the assumption on the of input choices. Capital is dynamic and quasi-fixed, determined at one-period ahead, $t-1$, and evolves in accordance with its law motion. Whereas the static or variable choices of labor and intermediate materials are made at time t. Therefore, at the beginning of each period, the information set $\mathcal{I}_t$ includes the current capital stock as a state variable. A-D.3 is very often seen in

the proxy variable literature, which requires either intermediate demand function to be strictly monotonic in the scalar productivity shock. The monotonicity can be guaranteed by imposing shape restrictions on the production and cost functions. Admittedly, A-D.3 is strong in the sense that one has to restrict the unobservables to be unit-dimensional, excluding market demand disturbances and input price differentials. To evoke A-D.3, an implicit assumption of homogeneous input prices need to be in place and its time variation is captured by time-varying functional form, provided it is common for all firms. Our identification strategy views A-D.3 as essential for inverting $\mathbb{M}_t(\cdot)$[8]. Another issue pointed by Ackerberg et al. [2] is the functional dependency problem. It says that labor would have no variation after the capital and materials inputs are controlled once the choice of labor input depends only on the state variable (ω_t, K_t). In their paper, they discuss additional conditions to resolve the functional dependency problem.

A-D.4 specifies the productivity evolutionary process unique in the dynamic model. It outlines the dynamic rule for the persistent unobservable to be Markovian, meaning that the distribution of the productivity at t given the past information set can be summarized all by the previous productivity at $t-1$. The second part of A-D.4 restricts the full independence of ε_t with all available information at t. The distributional assumptions in A-D.4 are weaker than commonly assumed AR(1), e.g. $\omega_t = \rho\omega_{t-1} + \eta_t$ or additive structure, e.g. $\omega_t = g(\omega_{t-1}) + \eta_t$, in previous literature. It would also be possible to augment the conditioning set with the capital stock, e.g. $P(\omega_t|\mathcal{I}_{t-1}) = P(\omega_t|K_t, \omega_{t-1})$, to capture more specifications. A-D.4 implicitly assumes the stationarity of the process, which is not necessary for identification. Furthermore, A-D.4 ii). is a little restrictive to the extent that η_t is assumed to be a scalar. Proposition 2.2 is essentially an identification result of $P_t(\omega_t|, \omega_{t-1})$. In particular, the conditional distribution of productivity can be identified and estimated by the conditional distribution of material input given the current information set. A-D.5 is the initial condition. It stipulates that the initial productivity is some monotonic deterministic transformation of the initial capital stock.

Proposition 2.2. *Let* $V_t = F_t(M_t|\mathcal{I}_{t-1})$ *where* $\mathcal{I}_{t-1} = (K_t, \{Z_{t-1}\}_2^t)'$. *Under Assumption*

[8]There has been work that attempt to relax the scalar unobservables.

A-D.2 to A-D.4, then $V_t = P_t(\omega_t|\omega_{t-1})$.

To see this,

$$\begin{aligned}\Pr(M_t \leqslant m|\mathcal{I}_{t-1}) &= \Pr(\omega_t \leqslant \mathbb{M}_t^{-1}(K_t, m)|K_t, \{Z_{t-1}\}_1^T) \\ &= \Pr(\omega_t \leqslant \mathbb{M}_t^{-1}(K_t, m)|\omega_{t-1}) \\ &= \Pr(\omega_t \leqslant w|\omega_{t-1}) = P_t(w|\omega_{t-1})\end{aligned}$$

The first equality applies the inversion of the intermediate input demand function. The second equality holds due to the Markovian property in A-I.2. And note that $V_t \sim U[0,1]$.

Proposition 2.3. *Under Assumption A-D.1 to A-D.4, then* $(K_t, L_t) \perp (\omega_t, \varepsilon_t)|V_1^t$ *where* $V_1^t \equiv (V_t, V_{t-1}, \cdots, V_1)$.

A-D.4 also stipulates that $(K_t, L_t) \perp \{\varepsilon_t\}_1^T$ for each t, indicating strict exogeneity of $\{\varepsilon_t\}_1^T$. So next I can only focus on controlling for the persistent productivity shock, ω_t. With loss of generality, suppose the exogenous productivity shock follows the evolution process $\omega_t = \mathbb{W}_t(\omega_{t-1}, \eta_t)$, where $\eta_t \perp \mathcal{I}_{t-1}$ for each t. Conditioning on $P(\omega_t|\omega_{t-1})$ is equivalent to conditioning on $F_\eta(\eta_t)$. To see this,

$$\begin{aligned}P(\omega_t|\omega_{t-1}) &= \Pr(\mathbb{W}(\omega_{t-1}, \eta) \leqslant \omega_t|\omega_{t-1}) \\ &= \Pr(\eta \leqslant \mathbb{W}^{-1}(\omega_{t-1}, \omega_t)|\omega_{t-1}) \\ &= \Pr(\eta \leqslant \eta_t)\end{aligned}$$

Since conditioning on $F_\eta(\eta_t)$ contains the same information as conditioning on η_t. Note that the conditioning set $(V_t, \mathcal{I}_{t-1})$ is not the same as $\mathcal{I}_{t-1}$ in that the former is richer by incorporating the variation of M_t given the information at $t-1$, which is not contained in $\mathcal{I}_{t-1}$. It is trivial that K_t is independent of ω_t since $K_t \in \mathcal{I}_{t-1}$. For L_t, it is also true because conditioning on η_t as well as ω_{t-1} completely pins down ω_t and no variation is left so $L_t \perp \omega_t|(\omega_{t-1}, \mathcal{I}_{t-1})$. If the true process is augmented by K_t, such as $\omega_t = \mathbb{W}_t(K_t, \omega_{t-1}, \eta_t)$, Proposition 2.2 still holds by augmenting V_1^t to $\{K_t, K_{t-1}, \cdots, K_1, V_1^t\}$.

One can also view Proposition 2.3 to be a dimension-reduction technique since

conditioning on the whole set of $\mathcal{I}_{t-1}$ is redundant in our setting. Note that the Markovian process implies $P(\omega_{t-1}) = P(\omega_{t-1}|\omega_{t-2})\cdots P(\omega_1|\omega_0)P(\omega_0)$. So by repeatedly using Proposition 2.2, i.e. for $\{t\}_2^T$, $P(\omega_t|\omega_{t-1}) = F_t(M_t|\mathcal{I}_{t-1})$. For $t = 1$, $P(\omega_1|\omega_0) = F_1(M_1|K_0)$, guaranteed by the initial condition assumed in A-I.5. Therefore $W_t \equiv \prod_{j=1}^{t} F_j(M_j|\mathcal{I}_{j-1})F_0(K_0)$ summarizes the distributional information contained in ω_{t-1} and thus can be used as one of the conditioning variable. Doing so will greatly reduce the dimension of control variables and have important implications in terms of nonparametric estimation. As for the estimation of V_1^T, several simplifications need to be made in order to reduce the dimension of the conditional information set. In Section 2.4, I suggest a semiparametric estimator together with variable selection.

Likewise for point identification of ASLPF and ASOE, a support condition in A-D.6 is needed. The support condition is to ensure that there is sufficient variation after conditioning on input choices. The validity has to be empirically examined.

A-D.6 Large support. $\text{supp}(V_1^t, |l, k) = \text{supp}(V_1^t) = [0, 1]$, for each t and each pair $(k, l) \in \mathcal{K} \times \mathcal{L}$.

Theorem 2.2. *Under Assumption A-D.1 to A-D.6, $\bar{f}_t(\cdot)$ (ASLPF), $\beta_{L,t}(\cdot)$ and $\beta_{K,t}(\cdot)$ (ASOE) are identified at each $(K, L) \in \mathcal{K} \times \mathcal{L}$ for each time period $t \in \{1, 2, \cdots, T\}$,*

$$\bar{f}_t(K, L) = \int E_t(y|K, L, v)dF_{V_1^t}(v) \tag{2.24}$$

$$\beta_{L,t}(K, L) = L\int \frac{\partial E_t(y|K, L, v)}{\partial L}dF_{V_1^t}(v) \tag{2.25}$$

$$\beta_{K,t}(K, L) = K\int \frac{\partial E_t(y|K, L, v)}{\partial K}dF_{V_1^t}(v) \tag{2.26}$$

The proof of Theorem 2.2 resembles that of the static model, with a static single control V_t replaced by the dynamic control vector, V_1^t. The final caveat: the above identification does not consider the entry and exit problems due to the selection on productivity. According to my empirical data and other research, not controlling for endogenous exits would cause downward biases of capital coefficients. To control for this, one can adapt the approach in Olley and Pakes [101] to the fully dynamic model considered in this paper.

2.4 Estimation and Testing

In this section, I present a fully nonparametric ASOE estimator for the static model. In principle, any nonparametric methods, such as local polynomial, sieve, and nearest neighbor estimators, can be used. Amongst those, I suggest the local linear estimator due to its convenience of estimating first order derivatives which are the building blocks to recover the average structural parameters defined previously. Moreover, along with the derivatives, the ASLPF can be estimated simultaneously. As the second objective of this paper is to propose an empirical valid test of Hick-neutrality productivity shocks, estimators of ASLPF therefore could be used to construct the test statistics. For dynamic models, I introduce a semiparametric variable selection estimator to reduce the dimensionality of conditioning set. Given the dynamic control obtained in the first stage, the second stage applies the local linear estimator resembling the static model.

2.4.1 Nonparametric Estimation of Static Models

The estimation can be capsuled in three stages. In the first stage, a per-period control covariate is estimated with the following local constant bias-reducing estimator. In the second stage, estimators of conditional expectation functions along with their first order derivatives are obtained by local linear estimation at each value. In the last step, the partial mean estimators of ASOE and ASLPF are obtained by averaging over the estimated control variate.

Stage 1: $\widehat{V}_{i,t}$ For each period, one can estimate the control variate by the bias-reduced local constant estimator like Eq. (2.27),

$$\widehat{V}_{i,t} = \widehat{F}_{M|K,t}(M_{i,t}, K_{i,t}) = \frac{\sum_{j=1}^{N_t} \mathcal{K}_h(K_{j,t} - K_{i,t})]\{\mathbf{1}[M_{j,t} \leqslant M_{i,t}] - \widehat{\delta}_{j,t}(K_{i,t})\}}{\sum_{j=1}^{N_t} \mathcal{K}_h(K_{j,t} - K_{i,t})} \tag{2.27}$$

where N_t is the sample size of period t and $\mathcal{K}_h(\cdot) \equiv \phi(\cdot/h)/h$ with h being the bandwidth and $\phi(\cdot)$ being the density function of standard normal. For convenience, the Silverman's *rule-of-thumb* bandwidth, e.g. $h = 1.06 \times std(K_t) \times N_t^{-r}$, where r is the window parameter. $\widehat{\delta}_{\cdot,t}(\cdot)$ is the difference of preliminary estimates between $V_{i,t}$ and $V_{j,t}$ to reduce the bias order

in the limit. See Shen and Klein [110] for details of the recursive bias reduction technique and the bandwidth selection. The conditional distribution, $\widehat{V}_{i,t}$ is essentially a conditional expectation of the indicator function given capital stocks. In the static model, it is necessary for the functional form of intermediate input demand to be varying over time in order to capture the change in the common factor prices.

Stage 2: In our empirical application, we suggest the local linear estimator for its convenience in estimating both the conditional mean and first order derivatives at the same time. In particular, one can solve the following weighted least square problem and the closed-form solutions are available. Let $U_t = (L_t, K_t, \widehat{V}_t)$. For each period t and each pair $u \equiv (L, K, V) \in \mathcal{L} \times \mathcal{K} \times [0, 1]$,

$$\begin{pmatrix} \widehat{C}_t(u) \\ \nabla \widehat{C}_t(u) \end{pmatrix} = \underset{c, \nabla c}{\arg\min}\, N_t^{-1} \sum_{j=1}^{N_t} \tau_{j,t} \left[y_{j,t} - c - (U_{j,t} - u)' \nabla c \right]^2 \mathcal{K}_h(U_{j,t} - u)$$

where $\tau_{j,t}$ denotes the estimated trimming function on U with fixed upper and lower bounds. It is well known that the closed form solution is available in Eq. (2.28)

$$\begin{pmatrix} \widehat{C}_t(u) \\ \nabla \widehat{C}_t(u) \end{pmatrix} = \left[\sum_{j=1}^{N_t} \tau_{j,t} \begin{pmatrix} 1 & (U_j - u)' \\ U_j - u & (U_j - u)(U_j - u)' \end{pmatrix} \mathcal{K}_{h,j}(u) \right]^{-1} \sum_{j=1}^{N_t} \tau_{j,t} \begin{pmatrix} 1 \\ U_j - u \end{pmatrix} y_{j,t} \mathcal{K}_{h,j}(u) \quad (2.28)$$

where $\mathcal{K}_{h,j}(u)$ is short for $\mathcal{K}_h(U_j - u)$ for brevity. Note that in the above formula, $\widehat{C}_t(u) = \widehat{E}(y_t | L_t = l, K_t = k, \widehat{V}_t = v)$ is the conditional expectation estimator—the most important building block to construct the test statistic below. For $\nabla \widehat{C}_t(u)$, it contains the partial derivative estimators of conditional expectation of log output with respect to input choices. Finally, the asymptotic properties of local linear estimators have been broadly studied in Fan and Li [34], Li and Racine [82]. So I will skip the discussion on how to obtain asymptotic variances as it can be found in many local polynomial literature.

Stage 3: The last stage delivers the ASOE estimators at each pair $(L, K) \in \mathcal{L} \times \mathcal{K}$ by the partial mean estimator with respect to $\widehat{V}$ evaluated at empirical points.

$$\widehat{\beta}_{L,t}(L,K) = N_t^{-1} \sum_{j=1}^{N_t} \nabla_L \widehat{E}_t(|L_t = L, K_t = K, \widehat{V}_t = \widehat{V}_{j,t}) \tag{2.29}$$

$$\widehat{\beta}_{K,t}(L,K) = N_t^{-1} \sum_{j=1}^{N_t} \nabla_K \widehat{E}_t(|L_t = L, K_t = K, \widehat{V}_t = \widehat{V}_{j,t}) \tag{2.30}$$

One can also take the marginal integration of $\nabla_L \widehat{C}_t(u)$ or $\nabla_K \widehat{C}_t(u)$ with respect to the estimated marginal distribution of $\widehat{V}_{i,t}$. But as the two methods are the same in the limit, either can be chosen. The average structural RTS estimator is naturally obtained by the sum, i.e. $\widehat{RTS} = \widehat{\beta}_{L,t} + \widehat{\beta}_{K,t}$. $\widehat{RTS}$ here can be different for firms with disparities in their input choices. A nice property of ASOEs is that it could capture more heterogeneity across firms with varying input combinations. However, in many cases, a highly summarized measure might be useful and can be easily compared to previous parametric estimators. Therefore, we suggest to take the empirical means of $\widehat{\beta}_{L,t}$ and $\widehat{\beta}_{K,t}$ to obtain the summarized average measures,

$$\bar{\widehat{\beta}}_{L,t} = N_t^{-1} \sum_{i=1}^{N_t} \widehat{\beta}_{L,t}(L_{i,t}, K_{i,t}) \tag{2.31}$$

$$\bar{\widehat{\beta}}_{K,t} = N_t^{-1} \sum_{i=1}^{N_t} \widehat{\beta}_{K,t}(L_{i,t}, K_{i,t}) \tag{2.32}$$

where $\widehat{\beta}_{L,t}(\cdot)$ and $\widehat{\beta}_{K,t}(\cdot)$ are local linear estimators of Eq. (2.24). Finally, a caveat is on the order of the bias: bias-reducing techniques or higher polynomials may have to be used in order to let bias vanish at $\sqrt{N}-$rate, which is not required for $\widehat{\beta}_{L,t}$ and $\widehat{\beta}_{K,t}$. In the empirical section, both ASOE and its average will be reported.

2.4.2 Semiparametric Estimation of Dynamic Models

The dynamic estimation resembles the procedure described in the static model. As alluded before, the problem of "'curse of dimensionality" could be detrimental even for moderate sample sizes. This is because the dynamic control variables $V_t = F_t(M_t|\mathcal{I}_t)$ is conditioned on the whole information set, $\mathcal{I}_t = \{K_t, (M_{t-1}, L_{t-1}, K_{t-1}), \cdots, (M_1, L_1, K_1)\}$ which could

become intractable as total time periods explode. Thus, this fact suggests the need for dimension reduction techniques. For practical purposes, I suggest to impose a single-index restriction on the information set under such case the conditioning set reduces to a single scalar, e.g. $I(\mathcal{I}_t, \alpha_0) = \mathcal{I}_t'\alpha_0$. For the semiparametric single index model, the finite parameters, α_0, have to be consistently estimated in the first stage. See Powell et al. [106], Ichimura [57], Klein and Spady [74], etc. for semiparametric index estimators. Furthermore, one can also perform the variable selection before the estimation once the dimension of $\mathcal{I}_t$ is too large compared with sample size. Variable selection has been widely studied in statistics, especially in the high-dimensional literature. Denote $\mathcal{I}_t^s$ to be the information set consisting of the selected and very important variables. Then under the required conditions, it is possible to reduce $F_t(M_t|\mathcal{I}_t)$ to $F_t(M_t|\alpha_0'\mathcal{I}_t^s)$. See Smith and Kohn [111], Fan and Li [33], Huang et al. [55], etc. for variable selection in nonparametric regressions.

2.4.3 Test Statistics of Hicks-neutrality

Now recall the hypotheses. As for the static model, I would conduct the test period-by-period. Doing so not only avoids the dynamic issues (such as entry-exit) but also provides an approach to examining the time effect. Under the null, the production function in period t in (2.1) exhibits Hicks-neutral technology as in (2.2).

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

An equivalent testable implication is obtained through the log transformation.

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

Following the approach in Gu [39], testing for Hicks-neutral productivity can be formulated as the nonparametric specification test of additive separability with multi-dimensional

unobservables. Likewise, I would use the empirical quantile mean (EQM) test proposed in that paper, based on the differences between average logged production functions under two competing specifications. Before conducting the inference, we need to replace the unknown control variable V_t by its bias-reducing estimator, $\widehat{V}_t$. By Theorem 7.2 in Gu [39], with the "generated" control variable, the asymptotic null distribution of the test statistic would not be affected. Then given the vector $\{y_t, (L_t, K_t), \widehat{V}_t\}$ for each t, corresponding to $\{Y, X, V\}$ in his setup, the test statistic and the consistent estimator of limiting variances can be thus constructed following the procedure. The details of the testing procedure will not be iterated in this paper.

2.5 Data

The empirical interests of this paper center on the U.S. manufacturing industries over 22 years, from 1990 to 2011. During this period, the contribution of the manufacturing industry to U.S. GDP has been declining from over 20% to 12% in 2011. Over the 22 years, the industry has experienced fast-growing in the 90s due to the rapid adoption of computer and electrical technology, which significantly drove the growth of the U.S. economy. Whereas after 2000, it has been hit by the Internet bubbles and the financial crises of 2008. Three striking features of the U.S. manufacturing industry are highlighted. Firstly, the output share of GDP has been stable over 50 years but since 2000, it has been gradually declining. After 2010, it has been surpassed by China measured by value-added. Secondly, the employment attributable to manufacturing has also been declining over time. Meanwhile, capital has become a relative more important factor. Thirdly, the productivity growth is imbalanced across sectors. Since the total industry consists of over 20 sectors, such as textile production, chemical, computer, etc., as noted by Baily and Bosworth [10], faster productivity growth has only been experienced by some sector such as computer production, while most others remain slow.

2.5.1 Data and Summary Statistics

I use an unbalanced panel of 5,088 manufacturing firms, 40,560 total observations, from Compustat North America fundamental annual database during the period 1990-2011. Compustat provides detailed firm-level financial and operative spreadsheet variables from which I construct output and input variables. I also supplement it with deflators and industry-level depreciation rates from Becker et al. [11], available from NBER website and industry-level annual average wages from Quarterly Census of Employment Wage (QCEW) collected by BLS. I divide sales and nominal values of inputs by their corresponding deflators, taken from Becker et al. [11], to obtain constant-dollar quantities.

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. 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. [11]. Following Olley and Pakes's method, I use the lagged investment when computing capital input.[9] 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.[10]

Firms use many inputs in their production, such as land, raw materials, electricity, labor, different types of capital, etc. To simplify the presentation of methods, I focus on the contribution only from two main inputs and hence estimate the value-added function. Table 2.1 provides some descriptive statistics of our full sample and selected subsamples. To make results manageable, I only present 5 representative sectors out of 21; each has a noticeable amount of presence in the whole sample. For output and input variables, each cell reports the average value. The average length of firm appearance in our sample is

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

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

around 12 years, reflecting the unbalanced nature of the panel data. I refrain from using the balanced panel to avoid the sample selection issue due to the endogenous liquidation decisions mentioned in Olley and Pakes [101]. Turning to the average value of input and output variables, it is obvious that there are large cross-sector differences as to the scale of production as well as input substitution patterns. Therefore, empirical results are for both all industry and sector-wise estimates and testing.

Another way to calculate firm's value added is as sales minus materials, deflated by the GDP price deflator. Sales is net sales from Compustat (SALE). Materials is measured as total expenses minus labor expenses. Total expense is approximated as Sales minus operating income before depreciation and amortization (OIBDP in Compustat). Labor expense is calculated by multiplying the number of employees from Compustat (EMP) by average wages from the Social Security Administration. The stock of labor is measured by the number of employees from Compustat (EMP). These steps lead to the value added definition that is approximated by operating income before depreciation and amortization plus labor expense.

Table 2.1: 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.

2.6 Empirical Results

2.6.1 Empirical Estimation Results

Average Output Elasticities The nonparametric estimates of average output elasticities are presented in Table 2.2. To eliminate dynamic problems such as entry and exit, I estimate the static firm-level production function year by year. I list three nonparametric estimators

according to different model specifications. For "NP-X" methods, all inputs are assumed to be purely exogenous and therefore usual nonparametric methods are used without controlling for endogeneity. Under such exogeneity, whether the production function is Hicks-neutral does not matter in terms of the estimates of average output elasticities. For "NP-Add.", we assume that labor is the endogenous variable and the production function is Hicks-neutral. The proxy variable approach, developed in the previous section, is used. For "NP-Nonsep.", labor is treated as endogenous in a non-Hicks neutral production function. This corresponds to the fully nonparametric case.

In Figure 2.2 and Figure 2.3, I plot the estimates of three specifications against year, by labor and capital respectively. Notice that there is a downward trend for labor coefficients and on the contrary, capital has become relatively important over time. It reflects the change of input substitution patterns in the production process. This partly explains why the employment attributable to manufacturing has shrank as motivated in the beginning. More importantly, from these two plots, it is not hard to see that output-labor elasticities are significantly overestimated without taking into account the endogenous choice. On the other hand, it results in very low estimates of capital coefficients. If the endogeneity is properly controlled, then labor elasticity estimates are reduced on average by 12.5% while it increases by only 9.4% for capital, indicating an upward bias on the return-to-scale parameter. Furthermore, increasing return-to-scale parameters can be always found assuming pure exogeneity. But once controlling for endogeneity, we can only observe the increasing or constant return-to-scale in the late 90s. Immediately after 2000, decreasing return-to-scale becomes more pronounced. This finding highlights the importance of controlling for endogenous inputs and is also in comply with the fact that the manufacturing industry has been declining since 2000. Now turn to the assumption of Hicks-neutrality. By comparing estimates of "NP-Add." and "NP-Nonsep.", I observe that the differences are very small in most of years. However, such discrepancies do persist all the time. For example, mistakenly imposing Hicks-neutral technology could result in an overestimate (underestimate) bias as high as 3.5% for labor elasticity (3.1% for capital elasticity).

		NP-X.		NP-Add.		NP-Nonsep.	
Year	N	$\widehat{\beta}_L$	$\widehat{\beta}_K$	$\widehat{\beta}_L$	$\widehat{\beta}_K$	$\widehat{\beta}_L$	$\widehat{\beta}_K$
1990	1818	0.779	0.242	0.674	0.323	0.657	0.343
		(0.018)	(0.014)	(0.017)	(0.014)	(0.034)	(0.028)
1991	1893	0.748	0.251	0.691	0.289	0.656	0.320
		(0.016)	(0.013)	(0.016)	(0.013)	(0.032)	(0.026)
1992	1997	0.764	0.249	0.654	0.330	0.664	0.328
		(0.017)	(0.014)	(0.017)	(0.014)	(0.034)	(0.028)
1993	2146	0.772	0.269	0.669	0.347	0.673	0.338
		(0.016)	(0.013)	(0.016)	(0.013)	(0.030)	(0.025)
1994	2219	0.789	0.240	0.652	0.344	0.668	0.331
		(0.015)	(0.012)	(0.015)	(0.012)	(0.029)	(0.023)
1995	2350	0.792	0.238	0.691	0.317	0.715	0.304
		(0.015)	(0.012)	(0.014)	(0.012)	(0.028)	(0.022)
1996	2419	0.743	0.281	0.669	0.338	0.672	0.345
		(0.013)	(0.011)	(0.013)	(0.010)	(0.025)	(0.020)
1997	2391	0.717	0.303	0.636	0.367	0.642	0.370
		(0.013)	(0.011)	(0.012)	(0.010)	(0.024)	(0.020)
1998	2251	0.738	0.312	0.699	0.343	0.706	0.340
		(0.015)	(0.012)	(0.014)	(0.012)	(0.028)	(0.023)
1999	2136	0.679	0.352	0.614	0.405	0.632	0.381
		(0.017)	(0.014)	(0.017)	(0.014)	(0.033)	(0.028)
2000	1975	0.565	0.443	0.494	0.498	0.489	0.492
		(0.021)	(0.018)	(0.020)	(0.017)	(0.039)	(0.033)
2001	1818	0.598	0.426	0.530	0.479	0.522	0.475
		(0.024)	(0.020)	(0.024)	(0.020)	(0.045)	(0.039)
2002	1766	0.567	0.462	0.483	0.531	0.463	0.543
		(0.026)	(0.022)	(0.025)	(0.022)	(0.049)	(0.042)
2003	1736	0.540	0.463	0.419	0.567	0.416	0.560
		(0.027)	(0.023)	(0.026)	(0.022)	(0.051)	(0.043)
2004	1683	0.578	0.413	0.430	0.529	0.436	0.508
		(0.029)	(0.024)	(0.028)	(0.023)	(0.054)	(0.045)
2005	1594	0.588	0.399	0.422	0.531	0.405	0.529
		(0.031)	(0.025)	(0.029)	(0.024)	(0.057)	(0.047)
2006	1527	0.638	0.357	0.506	0.462	0.436	0.488
		(0.033)	(0.027)	(0.032)	(0.027)	(0.061)	(0.051)
2007	1465	0.647	0.357	0.490	0.485	0.430	0.503
		(0.035)	(0.029)	(0.034)	(0.029)	(0.066)	(0.055)
2008	1345	0.643	0.365	0.422	0.542	0.359	0.549
		(0.038)	(0.032)	(0.037)	(0.032)	(0.073)	(0.062)
2009	1276	0.759	0.244	0.502	0.455	0.495	0.445
		(0.041)	(0.034)	(0.040)	(0.033)	(0.077)	(0.065)
2010	1278	0.783	0.235	0.553	0.427	0.515	0.449
		(0.040)	(0.034)	(0.039)	(0.033)	(0.075)	(0.064)
2011	1477	0.705	0.297	0.487	0.471	0.430	0.522
		(0.037)	(0.031)	(0.035)	(0.030)	(0.068)	(0.059)

Table 2.2: Empirical Estimation Results of Average Output Elasticities

Note: s.e. are given in parentheses. Smoothing parameters, $r_1 = 1/7, r_2 = 1/6$. Trimming parameters, $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$.

Figure 2.2: Average Output-Labor Elasticities

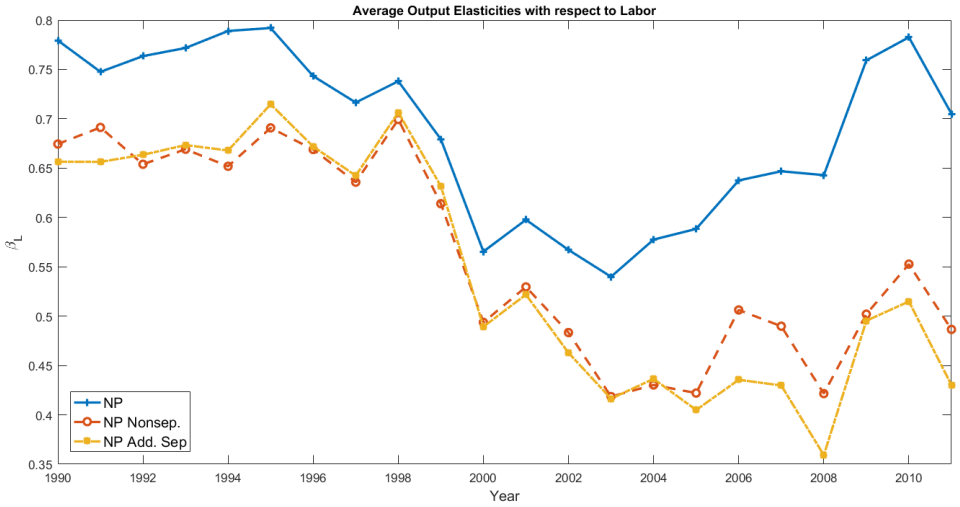


Figure 2.3: Average Output-Capital Elasticities

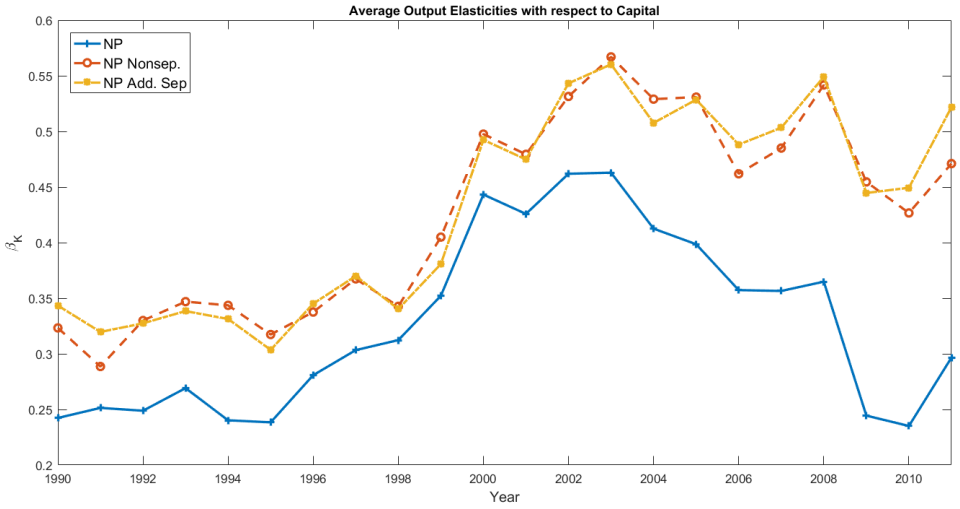


Table 2.3 presents average output-input elasticity estimates along with their standard errors for all manufacturing and selected industries. Four different models are compared. For comparison, I present the OLS estimates along with the nonparametric estimates of three different specifications. Firstly, note that on average, the U.S. manufacturing industry exhibits almost constant return-to-scale. Without taking the endogenous input choices, significant upward biases could be produced, no matter parametric or nonparametric. Secondly, huge heterogeneity is present across various sectors. The relative contribution

of factors could vary a lot and indicates heterogeneous technologies. For instance, sector 311, food product, are valuing labor and capital almost equally. In contrast, in fabricated metal product of sector 332, capital contributes over 80% in terms of the input shares.

Table 2.3: Empirical Estimation Results of Average Output Elasticities by Year

	OLS		NP		NP Add.		NP Nonsep.	
3-NAICS	$\widehat{\beta}_L$	$\widehat{\beta}_K$	$\widehat{\beta}_L$	$\widehat{\beta}_K$	$\widehat{\beta}_L$	$\widehat{\beta}_K$	$\widehat{\beta}_L$	$\widehat{\beta}_K$
All	0.688	0.322	0.693	0.324	0.577	0.416	0.567	0.418
	(0.006)	(0.005)	(0.008)	(0.007)	(0.008)	(0.005)	(0.015)	(0.013)
311	0.582	0.439	0.531	0.519	0.451	0.585	0.451	0.594
	(0.018)	(0.018)	(0.013)	(0.012)	(0.012)	(0.011)	(0.023)	(0.022)
325	0.805	0.259	0.823	0.249	0.735	0.317	0.761	0.295
	(0.010)	(0.008)	(0.011)	(0.009)	(0.011)	(0.009)	(0.020)	(0.016)
332	0.862	0.207	0.849	0.215	0.843	0.220	0.840	0.220
	(0.013)	(0.012)	(0.011)	(0.010)	(0.011)	(0.010)	(0.019)	(0.018)
333	0.839	0.197	0.850	0.200	0.776	0.259	0.790	0.253
	(0.011)	(0.009)	(0.010)	(0.009)	(0.010)	(0.008)	(0.019)	(0.016)
336	0.702	0.332	0.694	0.335	0.657	0.362	0.659	0.356
	(0.013)	(0.010)	(0.010)	(0.008)	(0.010)	(0.008)	(0.019)	(0.015)

Note: s.e. are given in parentheses. Smoothing parameters, $r_1 = 1/7, r_2 = 1/6$. Trimming parameters, $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$. OLS includes year dummies.

2.6.2 Empirical Testing Results

Table 2.4 provides empirical testing results. In this table, the test statistics and p-values of several tests for Hicks-neutrality are presented. To better summarize the findings, I plot our key testing results in the following figures.

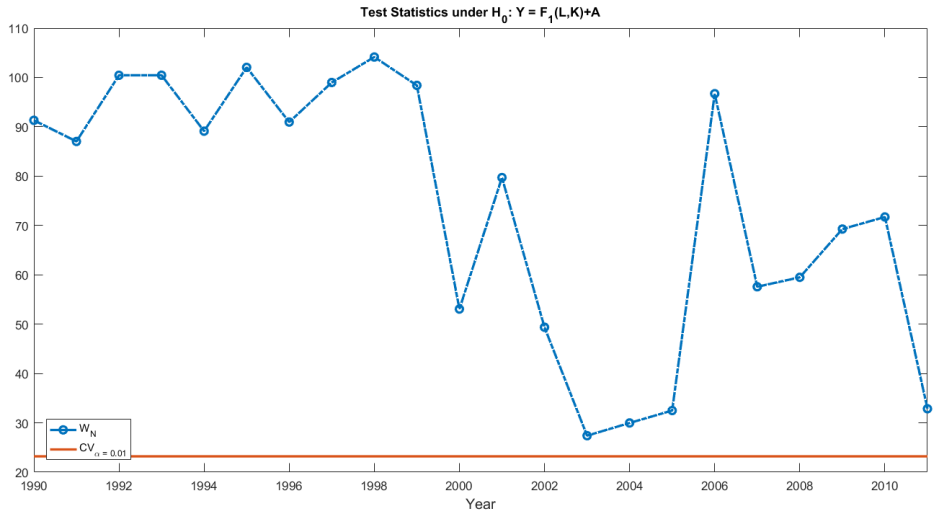
In Figure 2.4, I test the following hypotheses for each year,

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

The null hypothesis says that the production function is additive. This is a clearly false statement as many theoretical and empirical works have proved. I perform this test because I want to confirm the power of our proposed testing procedure in this empirical content. From the plot, I can reject 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 very good powers. This further gives us confidence in applying the test in more interesting scenarios.

Figure 2.4: Test Statistics by Year of Nonparametric Production Functions



In Figure 2.5, I present the testing results for the Hicks-neutrality for each year. As shown before, it is equivalent to testing the additive separability of the log-transformed production function,

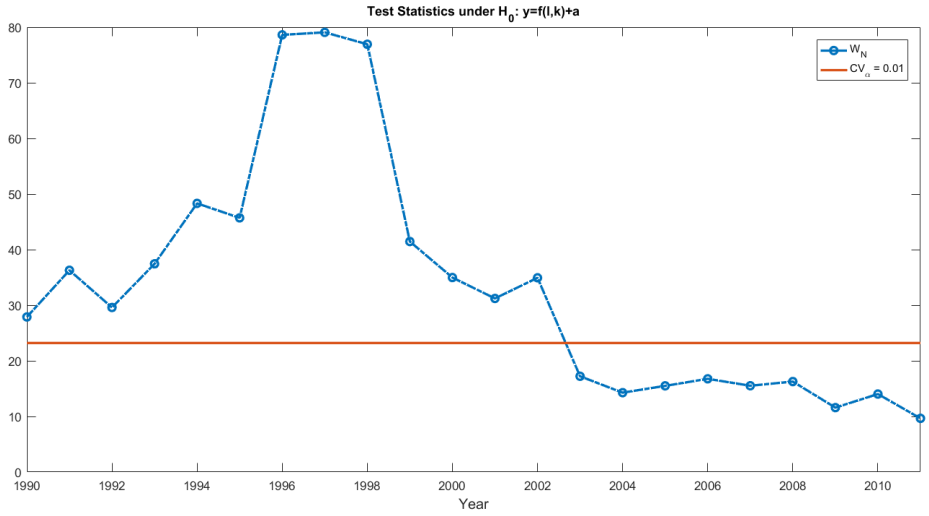
$$\mathbb{H}_0 \quad : \quad f_t(L,K,\omega,\varepsilon) = f_t^1(L,K) + a_t(\omega,\varepsilon), \quad \text{a.s.}$$

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

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, like 15%, 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 essential technologies start to converge again, as evidenced by the non-rejections of Hicks-neutral technological shocks.

Figure 2.5: Test Statistics by Year of Log Transformed Models



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, I do not expect a labor union strike to affect machinery and food product equally as substitution pattens of labor and capital in the production processes are quite different in those two sectors in terms of the relative shares of capital and labor, a point can be seen from Table 2.1. When there is sufficient large number of observations in each sector, one might mitigate this problem by placing sector dummy variables to control for this disparity. In light of this concern, I modify our test to include

sector-specific effects and derive a new set of testing hypotheses for each year,

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

where S represents sector dummies or sector specific effect. However, one is facing 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, only couple of observations are available in some years. Therefore, the variances of the test statistic could be extremely large and renders our test of almost no power. The limitation of data forces us to comprise on the full nonparametric sector-specific effects. To resolve this problem, we test with linear sector dummies, as commonly used in empirical works.

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

Under the null hypothesis, I assume the sector-specific effects enter in a linear way. 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 displayed in Figure 2.6. 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 may indicate that the essential technologies could be heterogeneous across firms even within the same sectors.

Figure 2.6: Test Statistics by Year of Log Transformed Models with Sector Dummies

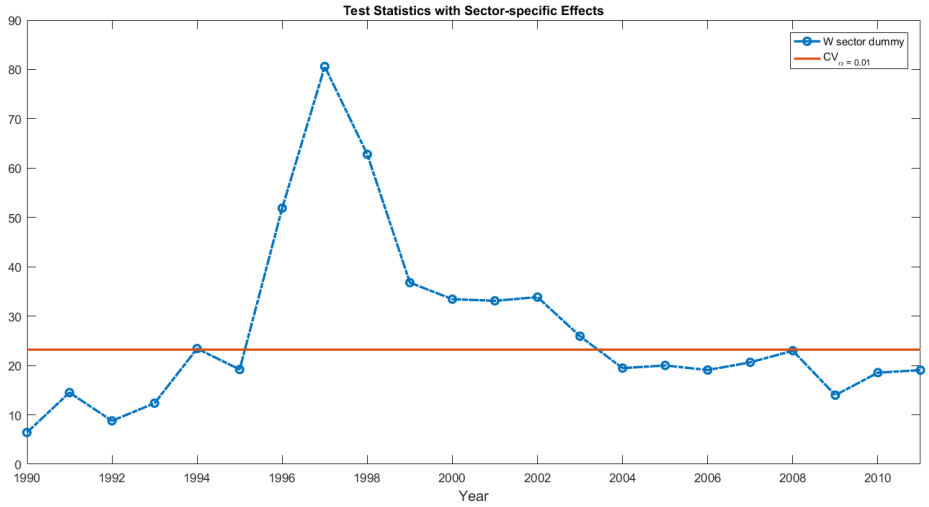


Table 2.4: Empirical Testing Results by Year 1990-2011

		Y		y		y^*	
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: Number of quantile $P = 10$. Smoothing parameters, $r_1 = 1/7, r_2 = 1/6$. Trimming parameters, $\kappa_1 = 0.01$ and $\kappa_2 = 0.025$.

2.7 Conclusions

In this paper, I consider the empirical implications of Hicks-neutral technological technology, a commonly employed assumption, on identification and estimation of nonparametric firm-level value-added production functions. Hicks-neutral technology puts significant restrictions on the substitution pattern of labor and capital in a production function. Without it, identification strategies need to change as they cannot be based on the log additivity of productivity shocks. Moreover, when wrongly imposed, the productivity measure might be severely distorted. I consider the identification and estimation of fully nonparametric firm-level production functions and empirically test the Hicks-neutral productivity in the U.S. manufacturing industry during the period from 1990 to 2011. In particular, I show that the *proxy* variable approach can be extended to fully nonparametric static and dynamic models with a set of slightly stronger conditions, in order to identify average structural output elasticities.

Secondly, I show that the Hicks-neutral restriction can be converted to the additive separability between inputs and unobservables in a monotonic transformed model for which the proposed testing procedure can be directly applied. With a panel data of the U.S. manufacturing industry, I find that there are periods in the 90s when the non-Hicks technological shocks occur which coincides with the mass adoption of computing technology. However, the productivity has thereafter become Hicks-neutral into the 2000s. Controlling for sector-specific effects mitigate this problem but not all of them. A conjecture is that firms have heterogeneous speed at adopting computer technologies and as a consequence create heterogeneity in input substitution patterns even within a sector. To confirm this, further research needs to be conducted.

.1 Proofs of Identification Results

Notation. $y_t = \ln Y_t, l_t = \ln L_t, k_t = \ln K_t, m_t = \ln M_t$ and $f_t(\cdot) = \ln F_t(\cdot)$.

Proof of Proposition 2.1. 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 $\omega = \mathbb{M}^{-1}(K, M)$

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

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

Proof of Theorem 2.1. The proof follows exactly the arguments of Theorem 1 in Imbens and Newey [60] by letting $X = (K, L)$ and $V = F_{M|K}(M, K)$. We permit the change of order of integration and differentiation to prove the identification results on average output elasticities. □

Chapter 3

Ordered Response Models with Unobserved Correlated Thresholds: An Application in Assessing Bond Overrating Bias

Jointly with Jiang, Yixiao and Yang, Shuyang

3.1 Introduction

Starting from the 1920s, credit rating agencies (CRAs) served the financial market by providing summarized information on the default risk of a security. The proper functioning of CRAs reduces information asymmetry between borrowers and lenders, and is crucial to the efficiency of financial market. Not only do individual investors rely on such uniformed rating scheme to assess of risk; legislators and regulators also use ratings as a benchmark to limit and regulate risky behavior of certain investors, such as insurance companies and pension funds. However, during the recent financial crisis in 2008, the mass defaults of highly rated structured financial products casts doubt on the accuracy and reliability of rating assessments from CRAs.

In this paper, we first propose a simple behavioral framework to model the bond rating process, accounting for the complex "liaison" between bond-issuing firms and the CRA. According to [76], the rating process can be decomposed into two steps. The CRA first constructs a default risk index using quantitative information from issuer's financial statements; the information set in this step is all publicly available. After obtaining a risk index, CRAs are likely to adjust the threshold points between rating categories based on their private "soft" information, so that bonds issued by two firms may end up with different ratings even though having the same observable characteristics. This so-called "soft adjustment" highlights the potential value of CRAs in a sense that they may utilize

their private information to affect the rating assignment when public information does not accurately capture the underlying default risk. Although the criteria for "soft adjustment" is vague and unspecified, common shareholders appear to be an important channel for conveying such private information. One implication of our model is that CRA-issuer relation can endogenously affect the issuing firm's decisions: having some private knowledge about the adjustment process will make firms alter their choices on bond characteristics. As a result, when modelling the rating process, variables that are strategically chosen by the firm cannot be treated as exogenous. Previous literature on bond rating has not explicitly addressed the endogeneity issue to our best knowledge. We highlight the importance of controlling for endogenous bond characteristics in the rating model and present empirical evidences of endogeneity arising from having the private information about firm-CRA liaison.

To empirically model the rating bias, we propose a semiparametric ordered response model with heterogeneous thresholds and endogenous regressors. The ordered rating is determined by placing a constructed latent default risk index in between a set of unknown pre-specified rating thresholds. In particular, we allow the set of thresholds to correlate with bond characteristics and be firm-specific to account for the individual-based soft adjustment process. Both the risk index parameters and thresholds across categories are our objects of interest. According to our behavioral framework, controlling for the issue-firm liaison is sufficient for eliminating endogeneity. In this paper, we provide two proxies for the *liaison*, namely, a observed constructed measure and an estimable control index. Then we rely on the control function approach to identify the index parameters. Such methods have been widely adopted in the literature of nonseparable models [16, 35, 59, 60, 65, 90, etc.]. For thresholds, we seek to identify and estimate the conditional mean thresholds relative to the base level, rather than the individual thresholds as the summarized measures suffice to answer our empirical question on bond rating bias. The identification strategy exploits the *conditional shift restrictions* between different categories, a generalization of Klein and Sherman [73] to models with endogenous regressors. We contribute to literature of location estimators studied in [88, 51, 78, 79, 73], etc.

We also provide a semiparametric two-stage estimator. In the first stage, the index

parameters are estimated by the weighted semiparametric least square (WSLS) estimator following the semiparametric literature [106, 58, 57, etc.]. In the second stage, we estimate the relative mean thresholds as a function of firm-CRA liaison measures by imposing shift restrictions between each adjacent categories. The consistency and asymptotic normality of our proposed estimators can be easily established and are also provided in the paper. To our best knowledge, our paper is the first that considers estimating thresholds in the presence of endogeneity.

After the crisis of subprime debt in 2008, the creditability of CRAs has attracted a lot of attentions. A main source of rating bias arises from the conflict of interests introduced by public ownership of CRAs. The largest agency, Moody's, became a publicly traded firm in 2001 while the second largest Standard&Poor's is owned by a public firm, McGraw-Hill Company. CRAs who are publicly traded have the incentives to be biased towards their shareholders, especially those who own dominant share of the agency. Several channels, through which such upward biases can occur, are examined and documented by past studies. First, public firms are operated under intensive pressure to grow and increase profits [17], which motivates CRAs to report inflated rating in order to retain repeated customers for rating fees under the current issuer-pay business model [27, 62]. Second and more importantly, CRAs' rating decisions can be directly influenced by the economic interest of their shareholders. Large shareholders tend to extract private benefit through governance power or threat of exit [3]. [66] further found that Moody's assigns favorable ratings to issuers related to its large shareholders relative to other CRAs.

Our data contains 5700 observations of individual bond rating history of 986 firms by Moody's Inc. from 2001 to 2008. We first compare the nonstructural and structural rating probability functions and confirm our observation: some bond and firm characteristics are indeed endogenous. To control for the endogeneity, we consider a single constructed investment share measure (termed MFOI) as well as a estimable relationship index to capture the omitted *liaison* in the rating criteria. Our estimation results on conditional thresholds suggest that the thresholds start to deviate from the baseline under impartiality, i.e. no CRA-issuer relation, indicating less strict criteria for assigning better ratings as connection strengthens. Moreover, overrating bias exhibits heterogeneous patterns across

rating categories. For grades A or above, the overrating bias starts to display only after the CRA-firm relation is stronger than the 70 percentile among the entire sample. On the other hand, in high-yield bond categories, even when CRA-issuer relation is as low as 20 percentile level, overrating bias starts to show up reflected by the decreases in average thresholds.

The layout of this paper is as follows. In Section 3.2, we study a simple behavioral model of bilateral bond rating that incorporates the interaction between issuers and CRA under private information. Section 3.3 formally defines the econometric model of ordered response and discusses conditions for identification. Section 3.4 proposes a two-stage semiparametric index and location estimator and derives its limiting properties. Asymptotic assumptions along with the proof are left in the appendix. Section 3.5 provides the institutional background, data and summary statistics. Empirical results are presented in Section 3.6. Finally, Section 3.7 concludes this paper.

3.2 A Simple Behavioral Model of Bond Ratings

CRAs use information on firm's financial statements and bond characteristics to assess credit risk, which reduces information-processing effort for investors. According to Moody's statement, it first utilizes quantitative information to estimate a default risk index according to a pre-specified financial metrics (FM), and then conducts various rating adjustments based on qualitative factors unobserved to investors. A bond will be placed into certain notches once its estimated default risk is in between the corresponding thresholds. Motivated by this fact, in this section, we try to understand the bilateral interaction between a single CRA and a representative bond-issuing firm and uncover the rating adjustment "blackbox" through a simple behavioral model consisting of three sequential stages. In the first stage, we assume that the CRA formulates an individual-specific rating criterion based on the *priori* CRA-issuer relationship through factors like common shareholders and business liaisons. The bond-issuing firm has a rational belief on the distribution of rating criterion given its private information on the *liaison*. In the second stage, a bond-issuing firm chooses bond characteristics by maximizing its expected capital gain given such belief. Finally, after an issuer independently chooses bond characteristics, an *ex-post* rating is reported to the public according to the CRA's internal rating models. In the rating process,

we want to feature two distinctions. First, we explicitly incorporate issuer heterogeneity by allowing rating thresholds to be stochastic and bond-specific. As a result, the unobserved threshold heterogeneity indicates that some bond and firm characteristics would become endogenous in CRAs' rating models.

3.2.1 Rating Agency: Rating Matrix and Bond-specific Thresholds

In this paper, we adopt and extend the threshold-crossing model that has been frequently employed in the literature of bond rating. Suppose that a linear latent default risk index can be computed from the CRA's rating metric, i.e. $Y_i^* = c_0 + X_i'\beta_0$, where $X_i = (X_i^F, X_i^B)'$ are a vector of firm and bond characteristics of bond i. In the rest of this paper, let $i \in \{1, 2, \cdots, N\}$ represent each bond (or single-bond firm) and $j \in \{0, 1, \cdots, J-1, J < \infty\}$ represent each rating notch or category. Now suppose the CRA places bond i into the jth notch in a threshold-crossing manner as below

$$Y_i = \sum_{j=0}^{J-1} j\{\widetilde{T}_{i,j-1} < Y_i^* \leqslant \widetilde{T}_{i,j}\} \tag{3.1}$$

where Y_i is the observed ordered rating with the support $\mathcal{Y} = \{0, 1, \cdots, J-1\}$. Also assume $\widetilde{T}_{i,-1} = \infty$ and $\widetilde{T}_{i,J} = \infty$. Let $\widetilde{\boldsymbol{T}}_{\boldsymbol{i}} = (\widetilde{T}_{i,0}, \cdots, \widetilde{T}_{i,J-1})$ denote a vector of unknown bond-specific stochastic thresholds. Under the above definition, those with $Y_i = 0$ are among the least risky bonds such as Aaa by Moody's standard. On the other hand, the larger Y_i, the riskier the bond.

Previous studies show that CRAs conduct "soft" adjustments beyond observed quantitative information on an individual bond basis. To model the adjustment process, we assume that the CRA could alter the bond-specific rating thresholds based on the "soft" information. Such adjustment enables the CRA to assign completely different ratings to bonds that have almost identical characteristics. It reflects the outcomes of qualitative assessment it has conducted on a specific firm or bond. For example, if the CRA has rated multiple bonds on a firm, it may develop in-depth knowledge about the off-balance-sheet risk management of this firm, which can then alter the rating thresholds. Moreover, CRAs may acquire private information about bond qualities through the liaison with common

shareholders. Specifically, the shareholders of the CRA, through investment in the stocks in bond-issuing firms or contracting with them in other businesses, form private connections that might be likely to produce upward (or downward) rating biases. Either way, the relation-based subjective adjustment conducted by the CRA would be reflected by the predetermined rating thresholds.

Now suppose that one can capture the liaison linking the two parties by some variable, R_i. The *ex post* private information, r_i is available to both the CRA and issuing firm i. We assume the set of firm-specific threshold $\widetilde{T}_{i,j}$ are additive separable functions of R_i and the idiosyncratic rating error $U_{i,j}$ which is irrelevant to the liaison.

$$\widetilde{T}_{i,j} = \widetilde{T}_j(R_i, u_i) = t_j(R_i) + U_{i,j}, \quad j = 0, \cdots J-1 \tag{3.2}$$

where we assume the full independence between $U_{i,j}$ and R_i. The additive separability is not essential as for the implications in this section but is a key assumption for identification in our empirical model. It indicates that each threshold can be decomposed into two additive terms, i.e. a category-varying component, $t_j(R_i)$ and a category-irrelevant one. The first component reflects that the heterogeneous "level effect" of a change of liaison on thresholds for each category. We argue that modeling the heterogeneous effects on thresholds is very crucial. For example, the CRA may respond quite differently to an increase of liaison for junk bonds versus investment-grade bonds. The second component in the additive model is invariant with categories, which can be interpreted as calculations errors of the CRA's agents or other exogenous adjustments.

Essentially, we need to allow the stochastic thresholds to be varying at bond-level rather than firm-level. For single-bond firms, there is no need to distinguish between these. However, for multi-bond firms, it is very likely that the liaison changes even between consecutive ratings for the same firm resulting from the quick shuffling of shareholder structures as such. Unless firms issue multiple bonds at the same time, which they seldom do. Therefore, in the empirical application, we choose to model the threshold heterogeneity at bond-level.

3.2.2 Firms: Contingent Choices of Bond Characteristics

Now consider the firm's issuing decisions. We assume that a bond-issuing firm chooses bond characteristics to maximize the discounted expected return on bond capital. We first consider the choice of issuing amount, and then the decision of subordination status.[1] It shall be pointed out later that both decisions are affected by the CRA-issuer liaison, R_i, through its influence on the rating rules.

To be specific, let $X^B = (X_1^B, X_2^B)$, where X_1^B denotes the issuing amount and X_2^B denotes the subordination status ($X_2^B = 1$ if the bond is senior; 0 otherwise). Since the liaison is common knowledge, bond-issuing firms may form a more accurate belief of the CRA's rating rule (or thresholds distribution) by conditioning on $R_i = r$, and then make financing decisions accordingly. For example, with a closer liaison, a higher expected rating may motivate the firm to issue more debt, declare subordination status more easily and undertake higher leverage ratios. For simplicity, it is also assumed that firms invest all what they can finance from the issuance in some businesses that on average pays the discounted return of ROI_i per dollar investment. We model a representative firm's financing decision of issuing amount as the following maximization problem given the liaison r and subordination status x_2^B,

$$x_1^B \quad = \quad \arg\max_{x \geqslant 0} E\left[ROI_i - C(Y_i, x_2^B)\right] | R_i = r \Big] x$$

where the expectation is taken with respect to the return of investment, ROI_i as well as the categorical rating, Y_i. $C(j, x_2^B)$ denotes the per dollar interest payments to investors for a bond with subordination status x_2^B and rating $Y_i = j$. The borrowing cost faced by the firm is largely determined by the CRA's reported rating. It can be conjectured that the interest cost is strictly decreasing with the ratings, e.g. $C(0, x_2^B) < \cdots < C(J-1, x_2^B)$. Given a particular rating $Y_i = j$, senior bonds are less costlier to finance, e.g. $C(j, 1) > C(j, 0)$. Firms make their issuing decisions based on the *ex ante* interest cost which can be written

[1] Among other characteristics that the firm may choose upon issuance, prior studies [103, 63, 13, .etc] have shown that issuing amount and subordination status are the two dominant factors that affect the borrowing cost.

as the average category-specific cost weighted by the rating distribution.

$$\begin{aligned} E[C(Y_i, x_2^B)|R_i = r] &= \sum_{j=0}^{J-1} C(j, x_2^B) \Pr(\widetilde{T}_{i,j-1} < c_0 + x'\beta_0 \leqslant \widetilde{T}_{i,j}|R_i = r) \\ &\equiv F(x_2^B, c_0 + x'\beta_0, \boldsymbol{t}(r)) \end{aligned}$$

While, as reflected above, the rating distribution is purely driven by the firm's default risk index, $y^* = c_0 + x'\beta_0$, and the set of firm i-specific thresholds, $\boldsymbol{t}(r) \equiv \{t_0(r), \cdots, t_{J-1}(r)\}$. By the second equality, define $F(\cdot)$ as some function determined by the joint distribution of U_j and the cost function $C(j, \cdot)$ for each j. Then given differentiability, the first order condition (FOC) with respect to the issuing amount x_1^B is

$$\overline{ROI} = F(x_2^B, c_0 + x'\beta_0, \boldsymbol{t}(r)) + \frac{\partial F(x_2^B, c_0 + x'\beta_0, \boldsymbol{t}(r))}{\partial y^*} x_1^B \beta_{01}$$

where ROI_i is assumed to be mean independent of the liaison, i.e. $E(ROI_i|R_i) = \overline{ROI}$. The FOC implies that the optimal amount x_1^B should equate the average return of borrowing to the cost of borrowing that consists two components. The first term measures the direct marginal cost for financing one more dollar given the current issuing amount and other characteristics. The second term is the indirect effect due to the marginal increase of cost from the change in default risk index. And this marginal increase would apply to all existing issuance.

Note that $F(\cdot)$ is a potentially nonseparable function of the default risk index y^* and the conditional thresholds, $\boldsymbol{t}(r)$. So from the FOC, the optimal amount x_1^B will be an implicit function of $\boldsymbol{t}(r)$. Since econometricians cannot observe the liaison r_i nor can they separate the level effect $t_j(r_i)$ from the idiosyncratic error u_i, the dependency between x_1^B and $\boldsymbol{t}(r)$ will induce some bond and firm characteristics to be endogenous. In contrast, most previous bond rating models have taken financial variables as purely exogenous. The exogeneity can only be justified when there is no private information of the firm-specific thresholds and the common distribution of thresholds is known to every bond-issuing firm. Nevertheless, in the presence of the liaison-induced private information, some bond characteristics would inevitably become endogenous due to firms' heterogeneous beliefs of

the thresholds distributions.

A similar argument could be applied to the firm's binary choice of subordinate status, X_2^B. Loosely speaking, it will choose to declare subordinate status if the net-payoff from declaring outweighs not declaring. For simplicity, one may assume that firms make issuing choices in a sequential way—first declaring the subordination status and then determining the issuing amount. Then it would be natural that the subordination status is correlated with the conditional thresholds through its correlation with X_1^B. One might challenge the behavioral assumption of sequential choices but the same implication could be derived even for simultaneous choices of both subordination status and other bond characteristics with a few additional complications. In addition, some firm-level financial variables, such as leverage ratios, would also depend indirectly on the thresholds. For example, consider a firm with multiple bonds, the debt-to-asset ratio increases as more bond is issued. Since the optimal issuing amount is correlated with the individual thresholds through the CRA-issuer liaison, then so would be the leverage ratio.

3.2.3 CRA: Final Reporting

After the bond characteristics are determined, the last step in the rating model is rather mechanical. Given the firm's choices X_i, the CRA calculates the *ex-post* default risk measure Y_i^* and adopts the pre-specified bond-specific rating thresholds $t_j(r_i)$ for each j. Then an independent rating noise, u_i, is randomly drawn from its distribution and being added to the threshold level as in Eq. (3.2). The realized error captures idiosyncratic shocks, such as miscalculations by the agent of CRA, unpredicted macroeconomic shocks, etc. Finally, a categorical rating for the bond is generated according to Eq. (3.1) and released to all investors and the public.

In this paper, we want to convey two important perspectives through the simple thought experiment above. First, qualitative adjustment of bond rating could be attributed to the heterogeneous thresholds that are partially determined by the bond-issuer private liaison. Second, failing to taking into account the heterogeneous thresholds might lead to the ignorance of the fact that some of firm and bond characteristics are endogenous in the rating model.

3.3 Ordered Response with Unobserved Correlated Thresholds

In this section, we study the econometric properties of the semiparametric ordered response models employed in the bond rating. Assumptions and identification conditions are given for the default risk index parameters and conditional threshold locations. In particular, recall that the ordered response model in Eq. (3.1), where the latent default risk index Y^* is converted into the observed ordinal information Y by the following transformation.

$$Y = \sum_{j=0}^{J-1} j\{\widetilde{T}_{j-1} < Y^* \leqslant \widetilde{T}_j\},\ Y^* = c_0 + X'\beta_0$$

where $\beta_0 = (\beta_{00}, \beta_{01}, \cdots, \beta_{0d})'$ is the coefficient vector conformable to the $d-$dimensional bond and firm characteristics X (continuous and discrete). In what follows, we suppress the i.i.d. subscript i for brevity when the context is self-evident.

Firstly, we identify and estimate the index parameters, up to location and scale, following the existing semiparametric literature [106, 58, 57, 74, etc.]. Secondly, we provide an identification analysis of heterogeneous thresholds as these are important to examine the existence of rating biases. And we show that with the additive separable idiosyncratic error, relative conditional thresholds are identified up to location and scale. Thresholds or Location estimators, though not mainstream, have already received some attention in the discrete choice literature [88, 51, 78, 79, 80, 73, 23, 24, etc.]. The closest to ours is the semiparametric ordered response threshold estimators by Klein and Sherman [73]. Their model is restricted to exogenous regressors and constant thresholds. Whereas our model allows one or more endogenous regressors that potentially correlate with the heterogeneous thresholds. To achieve identification, we rely on the control function approach that is frequently employed in nonseparable models with endogeneity [16, 35, 59, 60, 65, 94, 90, 48]. Furthermore, our models can be applied to other contents involving ordered responses or choices, beyond the empirical application considered here.

3.3.1 Identification of Endogenous Ordered Response

As is standard in the semiparametric discrete choice literature, several normalizations need to imposed. First, we argue that the additive index constant c_0 could not be separated from the stochastic thresholds if no distributional assumptions are imposed.[2] Furthermore, the parameters β_0 of the linear single index latent models are only identified up to scale [74]. Now partition the vector of regressors, $X = (X_0, \widetilde{X}')'$ and $X_0 \in \mathbb{R}$. Without loss of generality, let X_0 to be the continuous variable that may correlate with the stochastic thresholds, $\boldsymbol{T}$. It should be noted here that our model can take up multiple endogenous variables, both discrete and continuous alike. Next, redefine the identifiable index parameters by the division with respect to that of a particular continuous regressor, i.e., $\theta_0 \equiv (\gamma_{10}/\beta_{00}, \cdots, \gamma_{d0}/\beta_{00})' \in \Theta$ where Θ represents the finite-dimensional parameter space. Denote the identifiable index after the location-scale transformation by $V_0 \equiv X_0 + \widetilde{X}'\theta_0$.[3] In this paper, we only work with the linear index for simplicity. However, our model can handle general known nonlinear indices given identification. As for the thresholds, we are permitted to redefine the normalized thresholds after location-scale transformation to be $T_j \equiv (\widetilde{T}_j - c_0)/\beta_{00}$ for each level j. Depending on the sign of β_{00}, the interpretation of V^* might vary.[4] Permitted, we would subsequently work with the normalized threshold-crossing ordered response model in Eq. (3.3) which conveys the same ordinal information as the original model (3.1),

$$Y = \sum_{j=0}^{J-1} j\{T_{j-1} < V_0 \leqslant T_j\} \tag{3.3}$$

Note that several functions of economic and policy interest are implied from the above ordered structure. We first define the non-structural conditional cumulative rating function

---

[2] Any additive index disturbance, ε, as in $Y^* = c_0 + X'\beta_0 + \varepsilon$, cannot be separated from the stochastic thresholds. So one can assume that the thresholds absorbs all orthogonal disturbances.

[3] Such linear combination can incorporate nonlinear functions of X by redefining $X_j^+ = g_j(X)$ and $V = \sum_{j=1}^{d+1} \beta_{0j} g_j(X)$.

[4] In our empirical application, β_{00} is negative as it represents the marginal impact of total asset in the default risk.

given $v \in \mathbb{R}$ in Eq. (3.4)

$$P_j^n(v) \equiv \Pr(Y \leqslant j | V_0 = v) = \int \{v \leqslant T_j\} dF_{T_j|V_0=v}(t), \quad j = 0, 1, \cdots, J \tag{3.4}$$

$P_j^n(\cdot)$ measures the probability of being rating equal or above notch j given the default risk. However, $P_j^n(\cdot)$ is non-structural as the marginal effects of changes in bond characteristics on the this probability are confounded with the effect from changes of conditional distribution functions. A more interesting object, however, would only capture the partial effect on the probabilities due to the change in V_0 while holding the thresholds distributions fixed. This effect is summarized by the structural cumulative conditional rating probability function in Eq. (3.5) given $V_0 = v$,

$$P_j^s(v) \equiv \Pr(v \leqslant T_j) = \int \{v \leqslant T_j\} dF_{T_j}(t), \quad j = 0, 1, \cdots, J \tag{3.5}$$

where $F_{T_j}(\cdot)$ is the CDF of T_j. $P_j^s(v)$ corresponds to the average structural functions considered in Blundell and Powell [16], Imbens and Newey [60]. In the example of bond rating, $P_j(v)$ calculates the probability of being rated less than or equal to notch j, holding the threshold distribution unchanged for some default risk v. For models with only exogenous variables, $P_j^s(v)$ and $P_j^n(v)$ coincide with each other but diverge under correlated thresholds.[5] In Section 3.6, the empirical evidences further corroborates our conjecture on the endogenous bond characteristics.

Moreover, for continuous bond characteristics like $\widetilde{X}_k$, marginal effects of interest, $ME_{j,k}$ are subsequently available as the derivatives with respect to v multiplied by some scaling factor,

$$ME_{j,k}(v) = \triangledown P_j^s(v)\theta_{k0}, \quad j = 0, 1 \cdots, J$$

where $\triangledown P_j^s(v) \equiv \partial P_j^s(v)/\partial v$ provided existence. If X_k is some discrete characteristic such as the subordination status, the average treatment effects are trivially obtained as the

[5]The structural probability function of being rated exactly at notch j given $V_0 = v$ can be obtained straightforwardly by $\Pr(T_{j-1} < v \leqslant T_j) = P_j^s(v) - P_{j-1}^s(v), \quad j = 0, 1 \cdots, J.$.

difference between $P_j^s(v+\theta_{k0})$ and $P_j^s(v)$ with θ_{k0} being the corresponding coefficient.

Our model differs from traditional semiparametric ordered choice models in two important aspects. First, we incorporate the unobserved heterogeneity by allowing the set of thresholds to be individual-specific. Secondly, endogeneity could arise due to the heterogeneous thresholds as shown in the behavioral model in Section 3.2. As a consequence, usual identification strategies would fail unless the dependency is properly taken care of. In this paper, we rely on the control function approach to handle endogenous variables that are correlated with structural thresholds, in the spirit of Blundell and Powell [16], Florens et al. [35], Imbens and Newey [60]. We begin by assuming that a d_R-dimensional vector of control variables, denoted by R, satisfying Assumptions A-I.1 and A-I.2 stated below, is available to us. R may contain both discrete and continuous variables. Later in this section, we propose a semiparametric control index approach to handle the "curse of dimensionality" for nonparametric high-dimensional R as an extension.

A-I.1 Conditional Independence. $X \perp \boldsymbol{T}|R$ (and X and $\boldsymbol{T}$ are not measurable with respect to σ-field generated by R).

Assumption A-I.1 and A-I.2 are standard in the control function literature. In particular, A-I.1 specifies that R is the only possible reason for the dependency between X and $\boldsymbol{T}$. Moreover, X and $\boldsymbol{T}$ cannot be deterministic functions of R simultaneously. There are many ways to obtain the control vector R in practice. For example, when estimating return to education, the IQ test score is often taken to rectify the omitted variable bias resulting from missing intellectual ability. In triangular simultaneous equations models, the error term from first stage regression can be estimated and taken as a control variable. In our empirical application of bond rating, the CRA-issuer liaison are sufficient to control for the endogeneity as can be seen from Section 3.2. According to our behavioral model, bond characteristics X depend on the thresholds only through $\boldsymbol{t}(R_i)$, so the correlation would not be a problem once we can control the liaison, R_i. We will return to discuss the choices of R_i later in this section.

We now define the conditional cumulative rating function for Y being less than or equal

to notch j. Given $(x, r) \in \mathcal{X} \times \mathcal{R}$, or equivalently $(v, r) \in \mathbb{R} \times \mathcal{R}$,

$$\Pr(Y \leqslant j | X = x, R = r) = \Pr(V_0 \leqslant T_j | V_0 = v, R = r) \equiv P_j(v, r) \tag{3.6}$$

The second equality follows the fact that $V_0 \perp T_j | R = r$ due to the conditional independence assumption. From now on, our identification and estimation of θ_0 would rely on the moment conditions, $M(\theta) = (M_0(\theta), M_2(\theta), \cdots, M_{J-1}(\theta))'$ where

$$M_j(\theta) = E\left[\{Y < j\} | V(\theta), R\right] - E\left[P_j(V(\theta_0), R)\right] \tag{3.7}$$

A-I.2 Index Identification. There is a unique $\theta_0 \in \Theta$ such that $M(\theta_0) = 0$, defined in Eq. (3.7). For $\theta \neq \widetilde{\theta} \in \Theta$, then $\Pr(V(X, \theta) - V(X, \widetilde{\theta}) \neq 0) = 0$, w.p.1

A-I.2 assumes the point identification of the index parameter θ by the moment condition. A special case considered in this paper is when the index is linear, $V(X, \theta) = X_0 + \widetilde{X}'\theta$, for which a sufficient condition is $\det(\widetilde{X}'\widetilde{X}) > 0$ with X_0 being a continuous variable.

A-I.3 Large Support. $\mathcal{R} = \mathcal{R}^v, \quad \forall v \in \mathbb{R}, \quad$ a.s. where $\mathcal{R} = \text{supp}(R), \mathcal{R}^v = \text{supp}(R | V_0 = v)$.

A-I.3 requires the conditional support of R to be the same as the unconditional support. This sufficient support condition is often invoked in the control function literature to obtain point identification results of average structural functions. We require A-I.3 only for Proposition 1 on the point identification of $P_j^s(\cdot)$. As for the index parameters and relative threshold differences, the large support condition is not necessary. Proposition 1 states that $P_j(v)$ in Eq. (3.5) can be identified if the large support condition is invoked. The identification is achieved by marginally integrating out R for each index value $v \in \mathbb{R}$. The argument is standard and we leave the proof of Proposition 3.1 in the appendix.

Proposition 3.1. *Under Assumption A-I.1 to A.I-3, $P_j^s(v)$ and $ME_j(v)$ are identified for each $v \in \mathbb{R}$ and $j \in \mathcal{Y}$.*

3.3.2 Conditional Shift Restrictions

It has also been known that the ordinal structure carries some hidden information which can be revealed through the so-called shift restrictions between adjacent categories. It means that the cumulative probability of Y being less than or equal to level j can be related to that with respect to level k by shifting the location of thresholds. There exists hidden cross-level restrictions in the ordered model like Eq. (3.3). In Klein and Sherman [73], they use those shift restrictions to estimate the relative scaled thresholds in semiparametric ordered response models. For identification purpose, they require two conditions to be met. One is the single index structure, the other being the independence between X and the error term.[6] As for this paper, we generalize the shift restrictions to allow endogenous regressors with correlated thresholds. We show that the conditional and unconditional mean of relative thresholds can be identified under some additive separable structure of thresholds. And both objects have important implications and should be of great interest, especially concerning the empirical questions studied here.

We begin by defining notations formally. Let $\Delta_{j,k}$ denote the expectation of the *relative thresholds* or *threshold differences* between level j and level k, $T_k - T_j$. Let $t_j(r) \equiv E(T_j|R = r)$ denote the conditional expectation of the threshold j given the control variable $R = r$. $\Delta_{j,k}(r) \equiv E(T_k - T_j|R = r)$. $\Delta(R) = (\Delta_{0,1}(R), \Delta_{1,2}(R), \cdots, \Delta_{J-1,J}(R))'$ denotes a vector of conditional relative thresholds between adjacent levels. A related parameter of interest, resembling the "treatment effect on the treated", is defined as $t_j^c(x) = E(T_j|X = x)$ for each x. Analogously, let $\Delta_{k,j}^c(x) = E(T_k - T_j|X = x)$. In what follows, we focus on the identification of the conditional relative thresholds, which measure the corresponding shifts of average thresholds when R changes independently. Two mild and justifiable assumptions need to be imposed on the thresholds.

A-I.4 Additive Separability. $T_j = t_j(R) + U$, where $E(U|R) = 0$ for each $j \in \mathcal{Y}$.

A-I.4 is basically stating that for each $R = r$, the category-specific threshold can be decomposed into a component varying with respect to j and a non-varying component, U. $t_j(R)$ reflects the category-specific component depending on R. U, on the other

[6] The shift restrictions can be explored to derive more efficient index parameter estimators.

hand, representing the idiosyncratic error unobserved by the bond-issuing firm nor the econometrician. Under additive separability, we have a concrete interpretation of $t_j(R)$ as the mean thresholds given a particular liaison. It should be noted that A-I.4 does not rule out the full dependence between U and R. Only mean independence is required for the purpose of interpretation.

Proposition 3.2 (Conditional Shift Restriction)**.** *Under Assumption A-I.1, A-I.2 and A-I.4, for each* $(x, r) \in \mathcal{X} \times \mathcal{R}$ *and* $v = x_0 + \widetilde{x}'\theta_0$*, then* $P_j(v, r) = P_k(v + \Delta_{j,k}(r), r)$*, for each* $j, k \in \mathcal{Y}$*.*

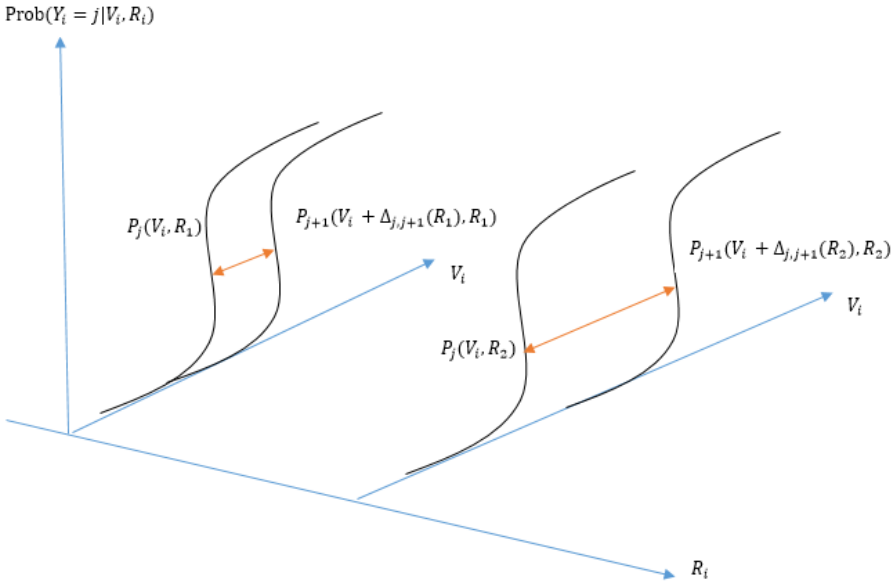
Proposition 3.2 states the conditional shift restrictions, a natural generalization of Klein and Sherman [73]. In particular, it reveals the hidden restrictions across categories. Figure 3.1 depicts the adjacent shift from the jth-conditional probability functions to the $(j+1)$th. Given the index $V_0 = v$ and the control variable $R = r$, one can equate $P_{j+1}(v, r)$ to $P_j(v, r)$ by simply shifting a constant distance of the index horizontally. The distance shifted is determined only by the value of control variable for a given pair of categories. By Assumption I.4, this distance can be interpreted as the relative conditional mean thresholds differences. This can also be seen from the following equalities. The last equality holds because the conditional distribution of U given R is invariant across categories by definition.

$$P_j(v, r) = \Pr(v \leqslant t_j(r) + U | R = r) = \Pr(v + \Delta_{j,k}(r) \leqslant t_k(r) + U | R = r) = P_k(v + \Delta_{j,k}(r), r)$$

We argue that A-I.4 is sufficient but not necessary to conduct shift restrictions. A weaker condition is the Assumption A-I.4 below. It states that the conditional distribution of the idiosyncratic disturbances are the same for each level j. For example, it includes some general nonseparable models such as $T_j = j + m(R, U)$, with m being some measurable function. However, we lose the interpretation of the $\Delta_{j,k}(r)$ being the relative mean thresholds differences given liaison measure $R = r$.

A-I.4' Distributional Invariance. $U_j | R \sim U_k | R$, for $j, k \in \mathcal{Y}$ where $U_j(R) = T_j - t_j(R)$, for each j.

Figure 3.1: Conditional Shift Restrictions from $P_j(V,R)$ and $P_{j+1}(V+\Delta,R)$



Proposition 3.3 (Identification)**.** *Under Assumption A-I.1, A-I.2 and A-I.4, for each* $r \in \mathcal{R}$ *and* $x \in \mathcal{X}$*, then* $\Delta(r)$ *and* $\Delta^c(x)$ *are point identified.*

In Proposition 3.3, we obtain identification results of relative conditional mean thresholds differences.The proof of Proposition 3.3 is straightforward given the invertability of $P_k(\cdot, r)$ for each r. For example, $\Delta_{j,k}(r) = P_k^{-1}(P_j(v,r),r) - v$, where $P_k^{-1}(P_k(v,r),r) = v$. Formal proofs of Proposition 3.2 and 3.3 are given in the appendix. Once the large support condition in A-I.2 can be justified, the unconditional mean of thresholds is subsequently obtained by taking the expectation, i.e. $\Delta_{j,k} = E[\Delta_{j,k}(R)]$. Once A-I.2 doesn't hold in some cases, the set-averaged expectations might exist. For instance, if one is interested in knowing the average difference on a compact set $\mathcal{R}^0 \subset \mathcal{R}$, then the set-averaged value can be obtained as $\Delta_{j,k}(\mathcal{R}^0) = E(\Delta_{j,k}(R)|R \in \mathcal{R}^0)$. The popular choices of R^0 include the quantile set, $\mathcal{R}^0 = \{r : Q_R(\tau_1) < r < Q_R(\tau_2)\}$ where (τ_1, τ_2) denotes respective lower and upper quantiles of interest and $Q_R(\tau) = F_R^{-1}(\tau)$ is the quantile function of R. For discrete R, $\mathcal{R}^0$ can be a countable subset of disjoint points.

The identification of $\Delta^c_{j,k}(\cdot)$ is established as follows,

$$\begin{aligned}\Delta^c_{j,k}(x) &= E(T_k - T_j|X = x) \\ &= E[E(T_k - T_j|X = x, R)|X = x] \\ &= E[\Delta_{j,k}(R)|X = x]\end{aligned}$$

The second equality uses the iterative expectation by firstly conditional on both X and R. The last equality holds because of A-I.1. Likewise, unconditional mean and conditional expectation over a set can be defined, though not being further pursued in this paper. In our application, we set the base level to be $Y = 0$, to which the relative thresholds are compared. So one can work with $\Delta^0(R) = A\Delta(R) = (\Delta_{0,1}(R), \Delta_{0,2}(R), \cdots, \Delta_{0,J-2}(R))'$ where A is a conformable lower triangular matrix with 1's below the diagonal.

Finally, we discuss the problems with recovering actual thresholds from the conditional relative differences. As the semiparametric ordered response model is only identified up to location and scale, recovery of true cutoff points is not possible without further assumptions even for constant threshold models. For the default risk index (or the creditworthiness), it is hard to pin down cutoff points when their lower and upper bounds are unknown. Even so, there are still very important implications that can be drawn from the relative thresholds estimates. As for Aaa bonds, it has been argued that no obvious biased threshold has been noted. On the contrary, significant overrating biases have been reported for bonds in lower notches. Our relative threshold estimators are able to detect and measure such deviations of bond rating biases. The counterfactuals would also shed light on public policies such as the reform of shareholder structures.

Besides, it has been noted that the ordered response model can be related to transformation models when Y has continuous meaning. Klein and Sherman [73] study the projected usage of a new telecommunication program in a transformation framework where the reported ordered usage Y is linked to the true usage Y^* by a monotonic transformation $Y^* = \Lambda(Y)$. They back out the true cutoff points by normalizing $\Lambda(0) = 0$ and $\Lambda(Y_{0.5}) = Y_{0.5}$, for which the reported usage is equal to the true usage at either 0 demand or median. Likewise, our models can also be related to a transformation model in the

same manner when there are clear meanings of the latent index. For example, in studying the effect of immigration on life happiness, the choice of immigration usually depends on individual-specific disposition towards the prospective workplace, living environment, etc. In health economics, when evaluationg the effect of certain medical treatment on self-reported health conditions, the selection of treatment might correlate with the patient specific thresholds.

3.3.3 The "Liaison" Controls

Controlling for the endogeneity is equivalent to controlling the CRA-issuer liaison in particular, as suggested by the behavioral model. However, it turns out that "liasison" is usually unavailable to empirical researchers. Moreover, even defining it properly is a rather hard task, let alone to quantify its relationship. Various proxy measures have been proposed in the bond rating literature. In this paper, we consider two different measures to control for the liaison. In the first case, we construct a single index of Moody-firm investment interaction (MFOI) by a weighted sum of the investment shares of common shareholders. The details of the index are presented in Section 3.5. We assume that MFOI is an appropriate proxy that is sufficient to control for the private information available to the bond-issuing firm. For simplicity, one can just let R be MFOI. In the second scenario, we assume that R is multi-dimensional. It could consist of observed information such as number of common shareholders, whether owned by influential investors, number of previous bonds rated, etc. in addition to MFOI. We will elaborate our choice of controls in the empirical section.

In cases where the control variables are relatively high-dimensional, we can consider using the control index to circumvent the "curse of dimensionality" in nonparametric models. To resolve this, we resort to the imposition of additional semiparametric index restriction on R as in Assumption A-I.5,

A-I.5 Control Index $\boldsymbol{T}|V_0, R \sim \boldsymbol{T}|V_0, L_0$, where $L_0 \equiv L(R, \alpha_0)$.

A-I.6 Index Identification There is a unique pair $(\theta_0, \alpha_0) \in \Theta \times \mathcal{A}$ such that $M_j(\theta_0, \alpha_0) =$

0 where $M_j(\cdot)$ is redefined in Eq. (3.8).

$$M_j(\theta) = E\left[\{Y < j\}|V(\theta), L(\alpha)\right] - E\left[P_j(V(\theta_0), L(\alpha_0))\right],\ j \in \mathcal{Y} \tag{3.8}$$

A-I.5 basically says that it is permitted to reduce the multi-dimensional R to a single linear index. Suppose that $R = (R_0, \widetilde{R}')'$ where $R_0 \in \mathbb{R}$ is a continuous variable for identification purpose, we assume a linear structure $L_0 \equiv R_0 + \widetilde{R}'\alpha_0$. In this case, the identification condition can be reduced to $|\det(\widetilde{R}\widetilde{R}')| > 0$ if there is no overlapping variables between X and R. Under A-I.5, the conditional cumulative probability function in Eq. (3.6) is reduced to a double-index model in Eq. (3.9)

$$P_j(v, l) = \Pr(Y \leqslant j|V(\theta_0) = v, L(\alpha_0) = l),\ \ j \in \mathcal{Y} \tag{3.9}$$

Such models have been studied in Ichimura and Lee [58], termed the semiparametric double index model. Identification assumption of (θ_0, α_0) requires the existence of at least one continuous variable in each index and a sufficient condition precludes the composition of same variables in both indices. For our model, these assumptions are automatically satisfied as we do not have the overlapping variables between indices. Proposition 3.4 is the direct result from Ichimura and Lee [58] on the identifiability of finite-dimensional parameters of multi-index models.

Proposition 3.4 (Identification of (θ_0, α_0))**.** *Under Assumption A-I.1, A-I.3 and A-I.5-A-I.6 , then* (θ_0, α_0) *are identified.*

3.4 A Two-stage Semiparametric Estimator

In this section, we provide a two-stage semiparametric estimators for $(\theta_0, \Delta(r))$ for each $r \in \mathcal{R}$ and give the asymptotic results in terms of consistency and asymptotic normality. In the first stage of estimation, we target at the single index parameters up to location and scale, θ_0, by WSLS with the moment conditions specified in the identification section and obtain the usual semiparametric estimator $\widehat{\theta}$ as well as a consistent index estimator $\widehat{V}_i = X_{i0} + X_i'\widehat{\theta}$. In the next stage, we estimate the relative conditional mean differences,

$\widehat{\Delta}(r)$ at each point $r \in \mathcal{R}$ by minimum distance estimators (MDE). Initial values at the second stage can be provided by the grid search method for $\Delta(\cdot)$ at each point in the support, which is attractive for its fast computing speed. Relevant estimators of interest can be subsequently obtained by numerical integrating out R on particular sets or summing over empirical points to obtain partial means over some regions or full sample. The two-stage estimator can be combined in a single-step GMM estimator with index moment conditions jointly with conditional shift restrictions, resulting in more efficient estimators, admittedly. Nethertheless, the two-step estimator would be much faster in practice especially when the dimension of parameter spaces tends to be large. One can postulate that our two-step estimator can significantly save the computational burden, especially when Y is divided into numerous level along with many control covariates. So we focus on the two-stage estimator in this paper and derive its large sample properties accordingly. A nice feature of the two-stage estimator is that the first stage estimation has no impact on the limiting distribution of the second stage as the converging speed of index estimators is faster than the conditional mean threshold function that converges at the nonparametric rate. In our application, we consider both cases of a single observable control variable as well as a partially estimable index. Due to the similarity, we illustrate our estimator and its asymptotic properties primarily with the single control case.

3.4.1 First Stage: Index Estimators

Conditional Probability Function We estimate the single index (or double index) parameters in the first stage by WSLS in the spirit of Ichimura and Lee [58], Ichimura [57]. In the presence of endogenous regressors, once the control variable is readily available, we could obtain consistent estimators by minimizing the (weighted) squared differences between the observed Y and its semiparametric conditional means. Various semiparametric single index estimators have been proposed. For exogenous covariates, see Manski [88], Powell et al. [106], Klein and Spady [74], Ahn et al. [4], Klein and Shen [69], etc. For models with endogeneity, see Blundell and Powell [16], Hoderlein and Sherman [48], etc.

First, we begin by introducing the estimator of conditional cumulative probability function of $Y_i \leqslant j$, denoted by $\widehat{P}_j(V_i(\theta), R_i)$. For any $\theta \in \Theta$, define $V_i \equiv V_i(\theta) = X_i + \widetilde{X}_i'\theta$

and we suppress θ for notational simplicity whenever it is self-evident. We use the Nadaraya-Watson kernel estimator to obtain the semiparametric conditional probabilities which will be used to construct the least square fitting. The leave-one-out semiparametric estimator of the conditional probability function for below or equal to category-j is specified in Eq. (3.10),

$$\widehat{P}_j(V_i(\theta), R_i) = \frac{\sum_{l\neq i}^{N} K_h(V_l(\theta) - V_i(\theta))K_h(R_l - R_i)\{Y_l \leqslant j\}}{\sum_{l\neq i}^{N} K_h(V_l(\theta) - V_i(\theta))K_h(R_l - R_i)} \tag{3.10}$$

Instead, for models of a control index with unknown parameters, define $L_i \equiv R_i + \widetilde{R}_i'\alpha_0$ and enlarge the parameter space to $\Theta \times \mathcal{A}$ to contain (θ_0, α_0).

As we seek for a $\sqrt{N}$-consistent parameter estimators, we resort to the bias reduction techniques available in the semiparametric literature to make sure that the asymptotic bias vanishes faster in the limit. In principle, one may use higher-order kernels, local smoothing and the recursive methods in Shen and Klein [110]. The performance comparison between various bias-reducing estimators is beyond the scope of this paper and is left for further research.

WSLS: First Stage Estimation of θ_0 We focus on the WSLS estimator proposed by Ichimura [57] to obtain a root-N consistent estimator of θ_0. WSLS can easily incorporate the moment conditions in Eq. (3.7) or Eq. (3.8) in the presence of the endogeneity arising from correlated thresholds. As the ordered response model is heteroskedastic in nature, the feasible WSLS estimator is used for efficiency concerns by solving the following minimization problem in two steps.

$$\widehat{\theta} = \underset{\theta\in\Theta}{\arg\min}\, N^{-1} \sum_{i=1}^{N}\sum_{j=0}^{J-1} \widehat{t}_i \left(\{Y_i \leqslant j\} - \widehat{P}_j(V_i(\theta), R_i)\right)^2 w_j(i;\theta_I) \tag{3.11}$$

or the double-index WSLS,

$$(\widehat{\theta}, \widehat{\alpha})' = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}}{\arg\min}\, N^{-1} \sum_{i=1}^{N}\sum_{j=0}^{J-1} \widehat{t}_i \left(\{Y_i \leqslant j\} - \widehat{P}_j(V_i(\theta), L_i(\alpha))\right)^2 w_j(i;\theta_I, \alpha_I) \tag{3.12}$$

where the weighting estimator is defined as $\widehat{w}_j() = \left[\widehat{P}_j(\cdot) - \widehat{P}_j(\cdot)^2\right]^{-1}$. θ_I or (θ_I, α_I) represents the consistent pilot estimator obtained from the first-step unweighted SLS. The

trimming function estimator $\widehat{t}_i = \Pi_{k=1}^{d_X+d_R}\{\widehat{q}_{Z_k}(\tau_l) < Z_{ki} < \widehat{q}_{Z_k}(\tau_u)\}$ is the product of the indicator functions for each continuous Z_k, with fixed lower and upper quantiles τ_l and τ_u, where $Z_i = (X_i', R_i')'$. $\widehat{q}_{Z_{ik}}(\tau)$ is estimated by the empirical quantile function, $\inf\{z_k : N^{-1}\sum_{i=1}^{N}\{Z_{ki} \leqslant z_k\} \geqslant \tau\}$. In the following, we also suppress the functional dependence on observable covariates, i.e. $\widehat{P}_j(i;\theta) \equiv \widehat{P}_j(V_i(\theta), R_i)$ to highlight the estimation problem.

Note that the two-step WSLS can be solved in an equivalent semiparametric pseudo-MLE framework similar to Klein and Vella [71], Maurer et al. [94]. It requires only one-step of optimization and can be computationally faster. Take the single index model as an example, define $\widehat{P}_{-1,i}(\theta) = 0$ and $\widehat{P}_{J,i}(\theta) = 1$.

$$\widehat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} N^{-1}\sum_{i=1}^{N}\sum_{j=0}^{J}\widehat{t}_i\{Y_i = j\}\ln\left(\widehat{P}_j(i;\theta) - \widehat{P}_{j-1}(i;\theta)\right)$$

Note that the solution, $\widehat{\theta}_{MLE}$, to the above optimization is identical to our WSLS estimator. We also need to point out that our estimator is not the most efficient one. One could obtain a more efficient estimator such as GMM by choosing the optimal weighting matrix across moment conditions or incorporating shift restrictions between levels. But we would not pursue them further here. As it can be seen later, this first-stage estimation variance would not affect the limiting distribution of conditional mean thresholds estimators in the second stage.

3.4.2 Second Stage: Conditional Mean Thresholds $\Delta(\cdot)$

We propose an extremum-type estimator that minimizes the distance between $P_j(V_i, r)$ and $P_k(V_i + \Delta_{k,j}(r), r)$ for each $r \in \mathcal{R}$, $j \neq k$ and $j, k \in \{0, 1, \cdots, J-1\}$, implied by the conditional shift restrictions. For a J-supported Y_i, there are totally $\binom{J-2}{2}$ possible restrictions to choose from. In order to have a parsimonious model, we only consider the shift conditions of adjacent levels. Additional restrictions could be used to increase the efficiency and perform overidentification test. Without redundant information, we are left with $J-2$ restrictions and for each $r \in \mathcal{R}$, the localized minimum distance estimator is

obtained by solving the following least square problem,

$$\widehat{\Delta}(r) = \arg\min_{\Delta} N^{-1} \sum_{i=1}^{N} \sum_{j=0}^{J-3} \widehat{\tau}_i(r) \left[\widehat{P}_j(\widehat{V}_i, r) - \widehat{P}_{j+1}(\widehat{V}_i + \Delta_{j+1,j}, r) \right]^2 \tag{3.13}$$

where $\widehat{V}_i = X_{0i} + X_i'\widehat{\theta}$ is obtained from the first stage and the trimming function estimator is given by

$$\widehat{\tau}_i(r) = \{\widehat{q}_{V|R}(\tau_l, r) \leqslant \widehat{V}_i \leqslant \widehat{q}_{V|R}(\tau_u, r)\} \tag{3.14}$$

where $\widehat{q}_{V|R}(,r)$ denotes the conditional quantile function estimator of V_i given $R_i = r$, estimated by inverting the smoothed conditional distribution function like $\widehat{q}_{V|R}(\tau, r) \equiv \inf\{v : N^{-1}\sum_{i=1}^{N} \widehat{F}_{V|R}(v|R = r) > \tau\}$ and $\widehat{F}$ defined in Eq. (3.15). (τ_l, τ_u) denotes the preset lower and upper quantiles.

$$\widehat{F}_{V|R}(v, r) \equiv \frac{\sum_{i=1}^{N} K_h(R_i - r)\Phi(\widehat{V}_i - v/h)}{\sum_{i=1}^{N} K_h(R_i - r)} \tag{3.15}$$

The relative conditional mean threshold of level j with respect to the base level (namely $Y_i = 0$) is readily available by multiplying a lower triangular matrix A with entry equal to 1 below and along the diagonal. Let $\widehat{\Delta}^0(r) = A\widehat{\Delta}(r)$, so $\widehat{\Delta}^0(r) = (\widehat{\Delta}_{1,0}(r), \widehat{\Delta}_{2,0}(r), \cdots, \widehat{\Delta}_{J-2,0}(r))'$. Localize-then-average estimator of conditional threshold difference over a measurable set $\mathcal{R}^0$ is computed like (3.16),

$$\widehat{\Delta}(\mathcal{R}^0) = \int_{\mathcal{R}^0} \widehat{\Delta}(r) d\lambda(r) \tag{3.16}$$

$\lambda(\cdot)$ is some measure[7]. $\lambda(\cdot)$ can also be taken as the CDF of R_i. For multivariate R_i, the empirical measure below is preferred.

$$\widehat{\Delta}(\mathcal{R}^0) = \frac{\sum_{i=1}^{N} \{R_i \in \mathcal{R}^0\} \widehat{\Delta}(R_i)}{\sum_{i=1}^{N} \{R_i \in \mathcal{R}^0\}} \tag{3.17}$$

[7]The integration of a vector should be taken as component-wise

For models where the control is given by an index estimated from the firs stage, $\widehat{L}_i = R_{0i} + \widetilde{R}_i'\widehat{\alpha}$, the above estimators can be modified by substituting $\widehat{L}_i$ for R_i. And one can also localize around the estimated index, $\widehat{L}_i$ if L_i can be interpreted as some CRA-issuer relationship measure.

In our empirical application, we estimate average effects by discretized weighted sum by the empirical difference of CDF of R_i, as shown in Eq. (3.18)

$$\widehat{\Delta}_{j,0} = \sum_{m_i=1}^{M} \widehat{\Delta}_{j,0}(R_{m_i})(\widehat{F}_R(R_{m_i}) - \widehat{F}_R(r_{m_{i-1}})) \tag{3.18}$$

where $m_i = \{1, \cdots, M\}$ that is the selected subset of the sample, instead of the whole sample. Doing so will increase the speed of computation, as generally $M << N$ when we have a very large data set. Simply replacing R_i with $\widehat{L}_i$ in Eq. (3.18) would give empirical estimators of average effects when multivariate control variables are present.

Finally, we consider the estimator of the structural and non-structural conditional probability functions defined in Eq. (3.4) and Eq. (3.5). Proposition 3.1 shows that $P_j^s(v)$ can be identified by integrating the conditional expectation function with respect to the CDF of R_i. Substitution with their consistent estimators gives us the estimators, $\widehat{P}_j^s(v)$, for each j. Likewise, the numerical integration can be simplified by the discretized sum as in Eq. (3.19).

$$\widehat{P}_j^s(v) = \sum_{m_i=1}^{M} \widehat{E}(\{Y_i \leqslant j\}|v, R_{m_i})(\widehat{F}_R(R_{m_i}) - \widehat{F}_R(R_{m_{i-1}})) \tag{3.19}$$

In contrast, the nonstructural conditional probability functions can be straightforwardly estimated as the conditional expectation function in Eq. (3.20) in which $\widehat{E}$ denotes the kernel estimator similar to Eq. (3.10).

$$\widehat{P}_j^n(v) = \widehat{E}(\{Y_i \leqslant j\}|\widehat{V}_i = v) \tag{3.20}$$

In Section 3.6, we plot $\widehat{P}_j^n(v)$ against $\widehat{P}_j^s(v)$ to empirically examine the endogeneity issue of bond characteristics.

3.4.3 Asymptotic Properties

In this section, we develop the asymptotic theory for the two-stage estimators of both $\widehat{\theta}$ and the relative thresholds estimator $\widehat{\Delta}(r)$ for each r in the support. In particular, Theorem 3.1 presents existing results on index parameter estimators and Theorem 3.2 gives consistency and asymptotic normality results on the conditional relative thresholds estimators. Finally, we obtain the consistent covariance matrix estimator by plugging in the estimators of each unknown component. To conserver space, asymptotic assumptions A-A.1 to A-A.6 are presented in the Appendix .2.1. Those assumptions are all standard in the non/semi-parametric literature. The proof of Theorems along with all supporting lemmas are left in Appendix .2.

Note that the consistency and asymptotic normality in Theorem 3.1 of the finite-dimensional index parameter estimators are very standard in the semiparametric literature. Since the first stage model reduces to the double-index model considered in Ichimura and Lee [58], Ichimura [57], we omit the proof. Theorem 3.2 a). shows that the relative thresholds estimators are pointwise consistent as the sample size increases; b). derives the asymptotic normality of the localized relative threshold estimators. A nice finding shows that the limiting variances of the relative thresholds estimators do not depend on the variability of the first-stage index estimators because the latter converge at a faster $\sqrt{N}$ rate than the nonparametric rate $\sqrt{Nh}$ for the second-stage estimators. As a result, not only could we obtain a simplistic form of the asymptotic variance of $\Delta(r)$, but also it permits us to analyze the variability separately from the index estimators.

Theorem 3.1 (Consistency and Asymptotic Normality of $\widehat{\theta}$)**.** *Under Assumption I.1-I.3 (and I.5, I.6 for estimable control index) and A.1-A.6, then as $N \to \infty$, it follows that*

$$a). \quad \widehat{\theta} \xrightarrow{\mathrm{P}} \theta_0$$

$$b). \quad \sqrt{N}(\widehat{\theta} - \theta_0) \sim N(0, \Omega^{-1})$$

where the covariance matrix is

$$\Omega \quad = \quad E\left[\sum_{j=0}^{J-1} t_i w_{j,i} \frac{\partial P_{j,i}(\theta)}{\partial \theta} \frac{\partial P_{j,i}(\theta)}{\partial \theta'}\right].$$

Theorem 3.2 (Asymptotic properties of $\widehat{\Delta}(r)$)**.** *Under Assumption I.1-I.4 and A.1-A.6, as $N \to \infty$, it follows that*

$$\begin{aligned} a). \quad & |\widehat{\Delta}_j(r) - \Delta_j(r)| = o_p(1),\ j = 0, 1 \cdots, J-2 \\ b). \quad & \sqrt{Nh_2}[\widehat{\Delta}(r) - \Delta(r)] \sim N(0, H(r)^{-1}\Sigma(r)H(r)^{-1}) \end{aligned}$$

where the covariance matrix is defined in Eq. (3.21) *and* (3.22)

$$\Sigma(r) \quad = \quad E[\xi_i(r)\xi_i(r)'] \tag{3.21}$$

$$H(r) \quad = \quad E\left[P_i'(r)P_i'(r)'\right] \tag{3.22}$$

where $\xi_i = (\xi_{0,i}, \xi_{1,i}, \cdots, \xi_{J-2,i})'$ *and* $P_i'(r) = \left(P_0'(V_i, r), P_1'(V_i, r), \cdots, P_{J-2}'(V_i, r)\right)'$*, in particular*

$$\xi_{j,i}(r) = \tau_i P_{j+1,i}'(r) K(R_i - r) \left\{ \frac{f(V_i)}{g(V_i, r)}[Y_{j,i} - P_j(V_i, r)] - \frac{f(V_i + \Delta(r))}{g(V_i + \Delta(r), r)}[Y_{j+1,i} - P_j(V_i, r)] \right\} \tag{3.23}$$

where $Y_{j,i} = \{Y_i \leqslant j\}$ *and the derivative of the conditional cumulative function with respect to* v

$$P_j'(V_i, r) = \frac{\partial P_j(V_i, r)}{\partial v} \tag{3.24}$$

A remark on bandwidth selection. As in Assumption A-A.5, bandwidths are allowed to be different for estimating $\widehat{\theta}$ and $\widehat{\Delta}(r)$. In order to eliminate the bias in Theorem 3.1, some bias-reducing techniques need to apply [69, 110, etc.]. However, to ensure the consistency of the Hessian matrix, it requires that the window parameter $r_1 < 1/8$ for double indices, ruling out the optimal bandwidth. In Theorem 3.2, bias reducing techniques still have to

be applied but it can be weaker than the first stage.[8]

Covariance Matrix Estimator We obtain the covariance matrix estimator by substituting with consistent estimators for each unknown component. For a given r in a compact set $\mathcal{R}_0$, the estimators of Σ_r are obtained in Eq. (3.25),

$$\widehat{\Sigma}(r) = \frac{1}{N}\sum_{i=1}^{N} \widehat{\xi}_i(r)\widehat{\xi}_i(r)' \tag{3.25}$$

where $\widehat{\xi}_i = (\widehat{\xi}_{0,i}, \widehat{\xi}_{1,i}, \cdots, \widehat{\xi}_{J-2,i})'$ and in particular

$$\widehat{\xi}_{j,i}(r) = \widehat{\tau}_i \widehat{P}'_{j+1,i} K(R_i - r) \left\{ \frac{\widehat{f}(\widehat{V}_i)}{\widehat{g}(\widehat{V}_i, r)}[Y_{j,i} - \widehat{P}_j(\widehat{V}_i, r)] - \frac{\widehat{f}(\widehat{V}_i + \widehat{\Delta}_j)}{\widehat{g}(\widehat{V}_i + \widehat{\Delta}_j, r)}[Y_{j+1,i} - \widehat{P}_j(\widehat{V}_i, r)] \right\} \tag{3.26}$$

For the Hessian matrix estimator,

$$\widehat{H}(r) = \frac{1}{N}\sum_{i=1}^{N} \widehat{P}'_i(r)\widehat{P}'_i(r)' \tag{3.27}$$

and in particular, $\widehat{P}'_i(r) = \left(\widehat{P}'_0(\widehat{V}_i, r), \widehat{P}'_1(\widehat{V}_i, r), \cdots, \widehat{P}'_{J-2}(\widehat{V}_i, r)\right)'$ where

$$\widehat{P}'_j(\widehat{V}_i, r) = \frac{\sum_{j\neq i}^{N} K'(\widehat{V}_j - \widehat{V}_i)K(R_i - r)\{Y_j \leqslant j\}}{h\sum_{j\neq i}^{N} K(\widehat{V}_j - \widehat{V}_i)K(R_i - r)} \tag{3.28}$$

$$- \frac{[\sum_{j\neq i}^{N} K(\widehat{V}_j - \widehat{V}_i)K(R_i - r)\{Y_j \leqslant j\}][\sum_{j\neq i}^{N} K'(\widehat{V}_j - \widehat{V}_i)K(R_i - r)]}{h[\sum_{j\neq i}^{N} K(\widehat{V}_j - \widehat{V}_i)K(R_i - r)]^2} \tag{3.29}$$

3.5 Bond Rating Industry and Data

3.5.1 Institutional Background

As the information intermediaries of the financial system, credit rating agency's primary function is to evaluate a particular debt instrument's (including government bonds, corporate bonds, CDs, etc.), credit worthiness utilizing their rating model and private information. By law or policy, some investors are only allowed to purchase bonds with an investment-grade ratings (Baa or higher). A corporation that can issue higher rated bonds

[8] For example, one can use third-order kernels in stage 1 and second-order kernels in stage 2.

has a lower borrowing costs compared to firms that cannot. Given that the ratings should reflect the true riskness of the bond, any material bias in rating agencies' decisions has the potential to impact the financial system and erode the market confidence.

In this study we argue that the increasingly public ownership of rating agencies might induce conflict of interests. As noticed by [21] and [119], the current credit rating industry is dominated by a few rating firms due to prudential regulation: with the "Big Three" credit rating agencies controlling approximately 95% of the ratings business. Moody's and Standard & Poor's (S&P) together control 80% of the global market, and Fitch Ratings controls a further 15% ([6]). Of the two biggest agencies Moody's became a public firm in 2001, while Standard& Poor's is part of the publicly traded McGraw-Hill Companies. Being a publicly traded firm not only intensifies the pressure to grow and increase profits ([17]), but also motivates the CRAs to be biased towards their own shareholders. As noted in [3], large shareholders may extract private benefit through their govenance power or threat of exit. For example, Warren Buffett, a major investor in Moody's, had to answer questions in front of the Financial Crisis Inquiry Commission in 2010[9] because media reports alleged that Moodys' has been slow to downgrade Wells Fargo, an investee of Berkshire Hathaway[10].

There are extensive studies on sources of rating bias. It is well known that the current issuer-pay model creates an incentive for the CRAs to assign inflated ratings. In previous studies, researchers ([62], [27]) have focused on compromised ratings on account of issuer-pay model. [75] and [76] argue that the CRAs might cater to borrowers with rating-based performance pricing agreements through their hard and soft adjustments. The paper that related to ours the most is [66], who documents that CRAs' rating decisions might possibly be affected by the economic interest of their large shareholders as well. They found Moodys' ratings on firms related with its large shareholders' are more favorable than S&P, and the difference cannot be explained by Moodys' private information.

In stead of examining whether the ratings are indeed affected by the CRA-issuer liaison through reduced-form regression, our model is structural in a sense that we identified the

[9] http://www.philstockworld.com/2011/03/14/transcript-of-warren-buffetts-testimony-in-front-of-the-fcic/

[10] http://www.forbes.com/sites/halahtouryalai/2012/02/16/missing-from-moodys-downgrade-list-warren-buffetts-favorite-bank/

pattern of overrating bias. As discussed in [76], after delivering a preliminary rating from the rating matrix, CRAs can adjust the ratings based on their private knowledge. Most naturally these adjustments can be reflected through a firm-specific set of threshold points: firms that are closely related with the rating agency ownership-wise might be assigned a less stringent thresholds so that it is easier for them to receive higher ratings. Once the issuer firms believe the CRA will give them favorable treatment, they might take more audaciously actions, e.g. issuing more debt, undertaking a higher leverage ratio, making some explanatory variables endogenous. This application focuses on those correlated cutoff points, which are structural object of interest.

In the aftermath of the 2008 crisis, there is a emerging literature focusing on regulation of credit rating agency. Different reforms/acts have been proposed to regulate the financial environment, and there are heated debate among them. In the famous Dodd-Frank Wall Street Reform and Consumer Protection Act (Pub.L. 111203[11], H.R. 4173[12]; commonly referred to and henceforce as "Dodd-Frank"), a entire section aims to improve the regulation of credit rating agencies. This law required the SEC to establish clear guidelines for determining which credit rating agencies qualify as Nationally Recognized Statistical Rating Organizations (NRSROs). It also gave the SEC the power to regulate NRSRO internal processes regarding record-keeping and how they guard against conflicts of interest[13]. See [102] and [119] for the importance of such oversight. The Franken-Wicker amendment to the Dodd-Frank financial reform law[14] would use a governmental entity to assign securities to qualified ratings agencies based on capacity and expertise.

3.5.2 Data and Summary Statistics

The data on the history of credit rating by Moody's is obtained from the Mergent's Fixed Income Securities Database(FISD). We exclude government bonds and retain all initial ratings on bonds issued by firms covered in both CRSP and Compustat, leaving us with a

[11] https://www.gpo.gov/fdsys/pkg/PLAW-111publ203/html/PLAW-111publ203.htm

[12] https://www.congress.gov/bill/111th-congress/house-bill/4173

[13] https://www.sec.gov/spotlight/dodd-frank/creditratingagencies.shtml

[14] https://www.sec.gov/comments/4-629/4629-28.pdf

final sample of 5700 new bonds issued by 986 firms from 2001-2008. Since this application features estimation of the overrating bias, we select the sample period after Moody's went IPO in 2000 and before publication of important regulation rules (for example, the Dodd-Frank in June 2009).

For each firm, short-term and long-term debt data are from quarterly Compustat-CRSP merged dataset. Short-term debt is estimated as the larger of Compustat items 118 ("Debt in current libabilities") and 224("Total current liability"). Long-term debt is taken from item 119("Total long-term liability"). The end of quarter stock price data and number of shares outstanding data are also taken from Compustat-CRSP.

Explanatory Variables When modeling the rating process, we follow previous literature in bond ratings to select firm/bond characteristics that determine the ratings (e.g, [103], [63], [13], [62], [20]). The explanatory variables are: (1) Issuer size, defined as the value of the firm's total asset (ASSET). (2) subordination status, a 0-1 dummy variable which is equal to one if the bond is a senior bond (SENIORITY). These two variable are claimed to be the most important ones in the regression framework ([63]). (3) Firm leverage, defined as the ratio of long-term debt to total assets (LEVERAGE). (4) Operating performance, defined as operating income before depreciation divided by sales (PROFIT). (5) Stability variable, defined as the variance of the firm's total asset in the last 16 quarters. (CVTA) (6) Issue size, defined as the par value of the bond issue (AMT). As motivated in the behavior model, LEVERAGE and AMT might be endogenous as firms might issue more debt when they "foresee" a chance of higher ratings.

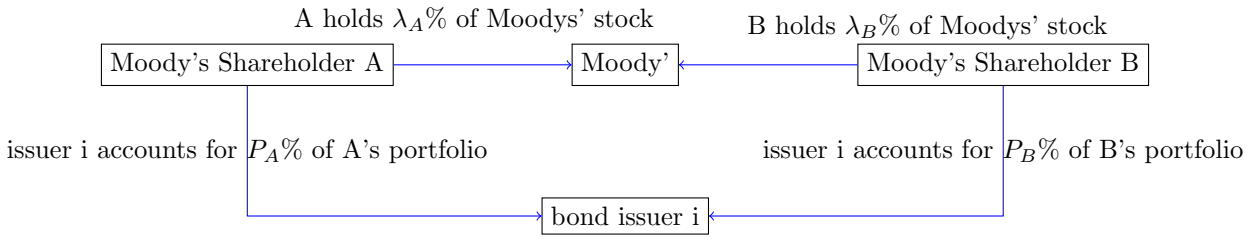
We take log on both sizing variables (AMT, ASSET) to make all covariates roughly have the same scale as their differences can be potentially very large. All financial ratios are computed using a 5-year arithmetic average of the annual ratios, as [63] points out that bond raters might look beyond a single year's data to avoid temporary anomalies. A summary statistics of the ratings and explanatory variables can be found in the upper panel of Table 3.1.

The Control Index The main variable that we use to control for endogeneity, termed Moody-Firm-Ownership-Interaction (MFOI), is defined as follow: Suppose an issuer firm i

Table 3.1: Summary Statistics

Variable	Mean	Std. Dev.	Min	Max
Rating and Rating Dummies				
Rating	3.953	1.404	1.000	7.000
Aaa	0.009	0.092	0.000	1.000
AA	0.156	0.363	0.000	1.000
A	0.246	0.431	0.000	1.000
Baa	0.268	0.443	0.000	1.000
Ba	0.131	0.337	0.000	1.000
B	0.164	0.370	0.000	1.000
C	0.027	0.161	0.000	1.000
Explanatory variables				
lnASSET	9.643	2.280	4.360	14.324
CVTA	0.230	0.169	0.003	1.416
LEVERAGE	0.264	0.178	0.002	1.212
PROFIT	0.026	0.058	-0.739	0.436
AMT	12.224	1.681	2.708	19.337
SENIORITY	0.809	0.393	0.000	1.000
Control Index				
numBOND	38.004	69.161	1.000	277.000
largeSH	0.606	0.706	0.000	3.000
numSH	160.116	114.214	0.000	419.000
MFOI	0.005	0.004	0.000	0.037

is jointly invested by two shareholders of Moodys, A and B.[15] The ownshership interaction between the issuer firm and Moodys' is presented in the following diagram:



We define bond issuer i's *ownership interaction* with Moody's as $MFOI_i = p_A\lambda_A + p_B\lambda_B$.

[15]However, Moodys could have shareholders who do not invest in the bond issuer i at all.

If Moodys' has M shareholders in total, we can generalize the above notion to:

$$MFOI_i = \sum_{j=1}^{M} p_j \lambda_j$$

by recognizing that $p_j = 0$ for shareholders who do not invest in bond issuer i. By construction, a bond issuer with larger $MFOI_i$ has a stronger interaction with Moody's through Moodys' shareholders. Since $\sum_{j=1}^{M} \lambda_j = 1$, $MFOI_i$ could be interpreted as issuer i's expected weight in Moody's shareholder's portfolio, which seems to capture the essence of the aforementioned CRA-issuer liaison. By conditioning on MFOI, issuers should no longer have the incentive to utilize their liaison with Moodys to obtain higher ratings.

In case this single control variable cannot fully control endogeneity, we also construct a "control index" by forming a linear combination of MFOI and some other measures listed below: (1) Number of Shareholder, defined as the number of common shareholders of Moodys and the issuer firm. (numSH). (2) Number of Large Shareholder, defined as the number of common shareholders of Moodys and the issuer firm who owns at least 5 % of Moodys' stock (largeSH). (3) Number of bonds, defined as the number of bonds issued by the firm that have been rated by Moodys (numBOND). Clearly the number of common (large) shareholders capture the CRA-issuer liasion. Compared to MFOI, the first two measure downplay a issuer firm's importance to Moodys' shareholders. numBOND is intended to capture the fact that Moodys might have a overrating bias towards their returning customers.

3.5.3 Discussion

To motivate that our selection of control function could imply the conditional independence assumption, we start by presenting some simple correlation analysis between the rating outcome and the control variables in Table 3.2. If the control varibles could indeed capture the effect of CRA-issuer liasion on ratings, we ought to see co-movement between the control variables and the rating outcome: issuers that are close to Moodys ownership-wise should be assigned higher ratings. This hypothesis is consistent with the findings in Table 3.2: for example, in the last column, when a firm's MFOI increases, their bonds are

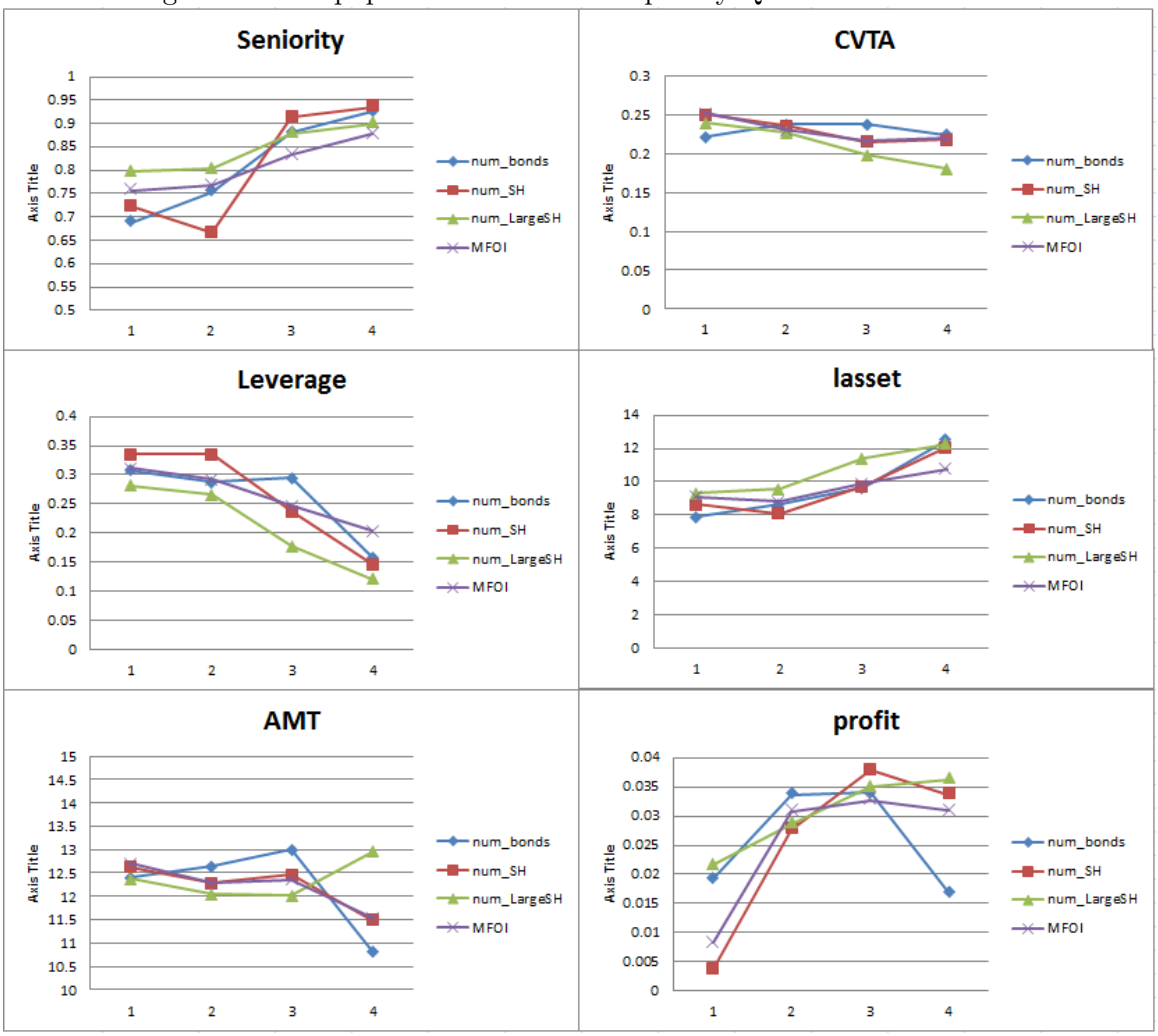
Table 3.2: Correlation between Control Variables and Rating Outcome

		numBOND	collectiveShare	largeSH	numSH	MFOI
Investment	Aaa	-0.0395	0.0186	0.0143	0.0198	-0.0115
Grade	AA	0.3725	0.285	0.3036	0.3907	0.2168
	A	0.2989	0.1275	-0.0568	0.2269	0.0348
	Baa	-0.2597	-0.0193	-0.035	-0.0697	-0.0234
High-yield	Ba	-0.1686	-0.116	-0.0713	-0.1918	-0.0622
	B	-0.2078	-0.2525	-0.1051	-0.3375	-0.127
	C	-0.072	-0.1187	-0.0541	-0.1311	-0.0893

more likely to be rated as AA or A and less likely to be rated below Baa. Similar patterns hold for other control variables. Through the correlation analysis, we also conjecture the effect of CRA-issuer relation on ratings might be fairly hetergeneous. As can be seen from Table 3.2, for bonds with extremely high or low ratings, the correlation between the rating and the control variables is quite small. Our model, in contrast with the standard linear probability or ordered probit model, could capture this heterogeneous effect by allowing the structural function P_k to be different in each category k.

Next, we motivate that our control variables are indeed correlated with some explanatory variables in the way predicted by the behavior model. Taken our main control, MFOI, for example, we report the subpopulation summary statistics grouped by different quantile level of MFOI in Table 3.3. Column 1-4 refers to the "group mean" and standard deviation for observations with MFOI in its 1-4 quantile. The average rating improves (recall that 1 indicates Aaa and 7 indicates C) as we move to a higher MFOI quantile, which is consistent with the findings in Table 3.2. In addition, we also notice significant trends in LEVERAGE and SENIORITY as we change the level of MFOI: when MFOI increases, firms issue less long-term debt, resulting a lower LEVERAGE; firms also declare a larger proportion of bonds to be SENIOR. These findings are consistent with our hypotheses that LEVERAGE and SENIORITY are endogenous from the behavior model. For other controls, we plot the subpopulation means of explanatory variables grouped by quantile levels of each control variable in Figure 3.2. It can be seen that the two hypothetical endogenous variables are indeed correlated with all proposed control variables.

Figure 3.2: Subpoplulation Means Grouped by Quantiles of Controls



Note: In each figure above, the x-axis corresponds to the quantile level of the control variable (numBOND, numSH, LargeSH and MFOI) and the y-axis corresponds to the mean level of the explanatory variable.

Table 3.3: Correlation between Control Variables and Rating Outcome

Variable	Q1 Mean	Std. Dev.	Q2 Mean	Std. Dev.	Q3 Mean	Std. Dev.	Q4 Mean	Std. Dev.
Ratings	4.269	1.475	4.334	1.262	3.828	1.331	3.382	1.329
Covariates								
lnASSET	9.101	2.238	8.807	1.812	9.892	2.145	10.774	2.363
CVTA	0.252	0.206	0.230	0.176	0.216	0.146	0.220	0.137
LEVERAGE	0.311	0.207	0.293	0.171	0.248	0.166	0.205	0.143
PROFIT	0.008	0.078	0.031	0.051	0.033	0.047	0.031	0.046
AMT	12.710	1.051	12.277	1.249	12.360	1.622	11.550	2.303
SENIORITY	0.756	0.429	0.767	0.423	0.833	0.373	0.878	0.328

3.6 Empirical Results

3.6.1 Index Coefficient Estimates

We begin by comparing index coefficient estimates obtained by the linear probability models (Linear thereafter), ordered probit (OProbit thereafter) and WSLS. The parametric and semiparametric models are estimated under three specifications and results are presented in Table 3.4. The first two columns present the estimates of linear probability models. It replicates the reduced-form model estimated by previous work (initially proposed by [53]), and show our control variables could capture additional variation of the data; therefore, intuitively we could control endogeneity by conditioning on them. We compare the results from linear probability models with and without the controls. Most coefficients from both models have the correct predicted sign: when a firm's PROFIT level increase, the probability of getting a higher rating on the firm's bond increases. When CVTA goes up (the variance of ASSET goes up) or the firm has a higher LEVERAGE ratio, the probability of getting a higher rating decreases. The issue amount (AMT) has a insignificant impact on ratings, but this finding is consistent with [66]. Moreover, a Likelihood-Ratio test indicates that the full model with controls has a significantly higher fit for the data at 1% confidence level (with a p-value equals 7.7e-9). Thus, we conclude our controls indeed capture some decent amount of unobserved variation. The next three columns show the results for the most commonly used OProbit model without controlling for CRA-issuer liaison, with a single control MFOI

as well as the control index constructed as the linear combination of R, respectively. The corresponding specification for WSLS are presented in the last three columns. For identification purpose, the coefficient of lnASSET is normalized to 1 and estimates are measuring the importance on the creditworthiness or negative default risk index relative to that of lnASSET. Similarly, for double index models of WSLS-M, the control coefficient of MOFI is also normalized to 1. All estimates of coefficients have the expected signs except for AMT which is around zero and small to some extent. Specifically, the asset volatility measured by CVTA has a negative effect on the creditworthiness, implying that the uncertainty in firm total assets may lead to higher probability of bond default. Profitability, as expected, is the most influential factor when determining the creditworthiness of bonds. Financial leverage, on the contrary, indicates the overall indebtedness status and a higher ratio reduces the credibility in remunerating its debt. Turning to the bond characteristics, we do find ambiguity in the offering amount. In general, as the estimates are around zero in small magnitude across specifications, as a matter of this fact, we postulate that its impact might be minimal in determining the bond default risk. As predicted by our rating model, declaration of seniority status adds safety insurance and reduces the associated default risk almost the same scale as lnASSET.

However, there are some significant differences between the parametric and semiparametric estimates concerning the magnitudes, despite the fact that they generate most of the same signs. Firstly, we note that WSLS estimates have enlarged relative effects of CVTA, PROFIT and SENIORITY on the negative default risk; while OProbit only captures the moderate effects. Since our WSLS is robust to misspecification of error distributions, this fact might indicate that assuming normal distributed random thresholds errors would lead to underestimated coefficients for some factors. Next, for models with multiple controls over CRA-issuer liaison, such as in column 3 and 6, the relative importances are significantly higher in WSLS than OProbit. Further, for our preferred method, WSLS, we find that whether or not controlling for liaison does not cause too much difference of index estimates which are somewhat robust to the selection of control covariates. The main takeaway regarding index estimators is that using parametric models such as OProbit may underestimate the relative importance of firm's profitability, asset

volatility and subordination status of the issued bonds in calculating the default risk index. So next, we focus on results of the preferred WSLS estimators.

Table 3.4: Estimation Results of First Stage Index Parameters

	Linear-X	Linear-M	OProbit-X	OProbit-S	Oprobit-M	WSLS-X	WSLS-S	WSLS-M
	$\widehat{\theta}$—Rating Index Parameter Estimates (with respect to LASSET)							
CVTA	-0.611	-0.585	-0.527	-0.534	-0.453	-1.124	-1.218	-1.199
LEVERAGE	-3.568	-3.503	-3.423	-3.424	-3.271	-2.980	-3.117	-3.162
PROFIT	15.570	15.275	15.383	15.429	14.719	26.816	26.630	28.595
AMT	0.027	-0.022	0.009	0.007	-0.042	-0.008	-0.017	0.148
SENIORITY	1.035	1.018	0.843	0.843	0.811	0.978	0.998	0.990
	$\widehat{\alpha}$—Control Index Parameter Estimates (with respect to MOFI)							
numSH		-0.016			-0.006			9.996
largeSH		-0.712			0.780			-0.884
numBOND		0.016			0.019			16.391
	$\widehat{\Delta}$—(Mean) Relative Thresholds Estimates (with respect to baselevel)							
$\widehat{\Delta}_{0,1}$			-3.378	-3.374	-3.294	-3.370	-3.447	-3.436
$\widehat{\Delta}_{0,2}$			-6.111	-6.100	-5.926	-5.636	-6.072	-6.959
$\widehat{\Delta}_{0,3}$			-8.632	-8.619	-8.380	-8.325	-8.749	-9.087
$\widehat{\Delta}_{0,4}$			-10.067	-10.053	-9.781	-9.944	-10.377	-10.441
$\widehat{\Delta}_{0,5}$			-13.352	-13.331	-12.965	-13.975	-14.481	-14.062

Note: 1. The suffix "-X" represents the exogenous case, "-S" the single control, "-M" the control index. 2. Estimates in the two upper panels reflect the relative importance on rating probabilities relative to the coefficient of lnASSET or MOFI. 3. Scaled thresholds estimates are computed relative to the base level of $Y = 0$, indicating the Aaa notch. 4. $\widehat{\Delta}$ for WSLS-S and WSLS-M are calculated by the discretized sum in Eq. (3.18) 5. The rule-of-thumb bandwidths, $h = 1.06std(R)N^{-r}$ are used, with the optimal rate i.e. $r = 1/6$ for double index models.

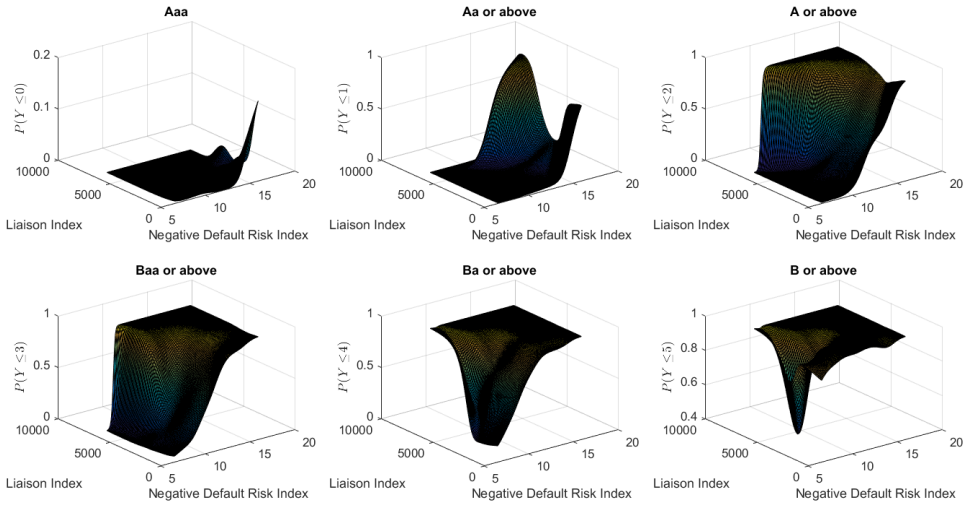
3.6.2 Conditional Probability Functions

Our model implies that when a firm issues new bonds based on the private information about the unobserved firm-specific thresholds through the investment liaisons with CRA, endogenous bond characteristics would naturally arise. A further implication is that the conditional probability of being rated into a particular grade would depend not only on the default risk index but also on the CRA-issuer liaison measures. If one believes that CRA would, on average, assign favorable ratings to those with whom a close relationship is shared, then it could be conjectured that the marginal effect of liaison on bond ratings is positive. Jiang (2016) calculate it using differences between conditional probability functions and confirm the existence of such effect. In this paper, we present and visualize the conditional cumulative rating probability functions in the 3-D figures. For the single control case, in Figure 3.4 , $x-$coordinate depicts the negative default risk index, $-X'\widehat{\theta}$, $y-$coordinate

depicts Liaison measure, MFOI and $z-$coordinate depicts the conditional probability of being rating in a particular grade or above. For the investment grade, Aaa, the common shareholder investment relationship has almost zero effect on the rating probability. Being rated into this category requires the default risk to be somewhat minimal, reflected by the almost flat hyperplane with z-coordinate close to 0. However, for the category of Aa or above shown in the second picture, we can notice the CRA-issuer relationship starts to work in the supposed direction. To be specific, given a certain level of creditworthiness, the probability of being rated Aa or above rises as the liaison becomes closer. Such pattern is not very obvious for A or above, Baa or above and Ba or above. Implicitly, this may indicate the fact that liaison has heterogeneous effects on rating probabilities at the notch level. The heterogeneity can be also seen from the last figure for category B or above. It looks that after a certain level of creditworthiness, any bond can be rated to be high-yield grade, or risky bond. The liaison only works in the predicted direction for those that contain substantive risks.

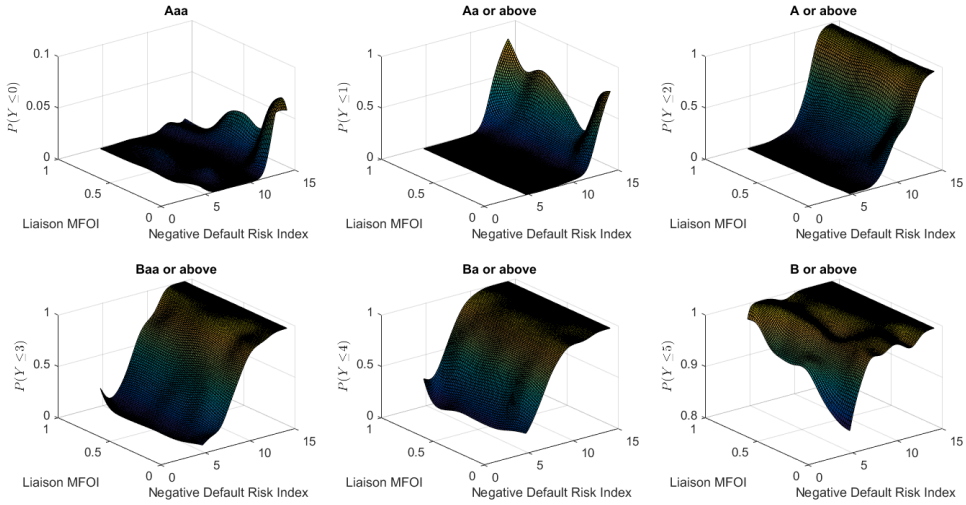
Figure 3.3 presents the 3-D visualization under the control index case in which an estimated linear index is used, instead of only the MFOI, in order to control other liaisons other than the investment relationship by common shareholders. The patterns shown in Figure 3.3 are roughly analogous to Figure 3.4, but with more apparent liaison effects. Starting from the second graph of Aa or above, it is obvious that given some reasonable level of default risk, the conditional probabilities have been driven up as liaison becomes tighter. Such effect continues as we consider the conditional probabilities of rating A or above, Baa or above and Ba or above. The last graph tells the same story as the single control case, when the default risk is around the thresholds of grade B, cultivating a good relationship with CRA can be very effective, at the margin, to level up those that would have been rated in the "junk" category.

Figure 3.3: Conditional Cumulative Rating Probability Functions $\widehat{P}(Y \leqslant j|\widehat{V}, \widehat{L})$



Note: 1. 100-by-100 Grid points are generated from $[\widehat{V}_{.05}, \widehat{V}_{.95}] \times [\widehat{L}_{.05}, \widehat{L}_{.95}]$ with equal interval, where subscripts represent the preset percentiles. 2. $\widehat{P}(Y \leqslant j|\widehat{V}, R)$ are estimated by the kernel estimator defined in Eq. (3.10) with rule-of-thumb bandwidth and optimal rates.

Figure 3.4: Conditional Cumulative Rating Probability Functions $\widehat{P}(Y \leqslant j|\widehat{V}, R)$



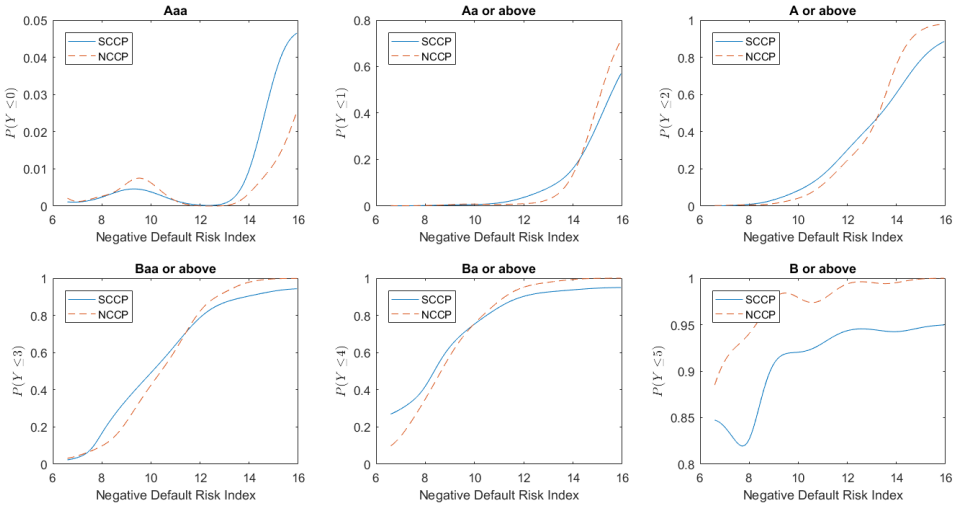
Note: 1. 100-by-100 Grid points are generated from $[\widehat{V}_{.05}, \widehat{V}_{.95}] \times [R_{.05}, R_{.95}]$ with equal interval, where subscripts represent the preset percentiles. 2. $\widehat{P}(Y \leqslant j|\widehat{V}, R)$ are estimated by the kernel estimator defined in Eq. (3.10) with rule-of-thumb bandwidth and optimal rates.

3.6.3 Empirical Evidences on Endogeneity

In many previous empirical applications, only the nonstructural cumulative conditional rating probability (NCCP) functions are calculated, like Eq. (3.4) under the premise that

all bond and firm characteristics are exogenous or invariant to the thresholds and error distribution. Through our structural rating model, we highlight the fact that the issuing amount and subordination status are a firm's particular choices given its belief on CRA's assigned cutoff points. While due to the impartiality through some CRA-issuer liaisons, these assigned ratings might be inflated or biased relative to their true grades. Under this scenario, a more useful object is the structural cumulative conditional probability (SCCP) function defined in Eq. (3.5), by which the rating probabilities and the marginal effects calculated could reflect the partial effects due to the change only from firm and bond characteristics without altering thresholds distributions. In Figure 3.5, we plot the SCCP versus NCCP for each category against the estimated control index. Generally speaking, gaps between the two curves become wider, as the negative default risk index gets larger in that NCCP tends to overestimate the rating probability function given default risk. As a result, marginal effects based on NCCP could not reflect the exogenous change of bond and firm characteristics alone, and in other words, are confounded with the indirect effect due to the change of conditional distributions of thresholds. This fact also provides evidence on the endogeneity of bond and firm characteristics as motivated in this paper, because otherwise one would not notice any difference between NCCP and SCCP given any level of default risk. This phenomenon displays heterogeneous patterns across rating categories as depicted in Figure 3.5. In contrast to SCCP, the NCCP, for rating between Ba and Aa, tends to be smaller when default risk is high and larger when default risk is low. The last graph indicates that there might be significant overrating bias if the endogeneity is not properly taken into account as the SCCP is uniformly below the NCCP over the whole range. In sum, if one does ignore the endogeneity of bond and firm characteristics, misleading counterfactual results are very likely to be produced. To construct a formal test of endogeneity is also possible.

Figure 3.5: Structural and Nonstructural Rating Probability Functions—Control Index



Note: 1. 100 Grid points are generated from $[\widehat{V}_{.05}, \widehat{V}_{.95}]$ with equal interval, where subscripts represent the preset percentiles. 2. SCCP and NCCP are calculated according to Eq. (3.19) and Eq. (3.20). 3. Kernel estimators with rule-of-thumb bandwidth and optimal rates are used.

3.6.4 Mean Thresholds

The third panel of Table 3.4 presents the average relative thresholds across various specifications. Recall that our relative thresholds are defined as $(T_j - T_0)$ for each j. Firstly, Neither OProbit with or without controls nor WSLS without controlling for random thresholds, is able to capture the heterogeneous relative thresholds by the modeling restrictions. Therefore we interpret their threshold estimates as the average exogenous thresholds in order to compare with ours computed as the discretized weighted sum in Eq. (3.18). The mean relative thresholds estimates are surprisingly robust to various parametric or semiparametric estimators without controlling for the liaison. In addition, those estimates are stable even for Oprobit with single or multiple controls. The relative differences, $\widehat{\Delta}_0$, are intuitive as can be seen that the cutoffs become less stringent as bonds are being rated into less favorable (and above) investment notches. For example for column 4, $\widehat{\Delta}_{0,1} = -3.370$ means that the minimum creditworthiness required, on average, for a bond to be rated in grade Aa or above if we normalize the mean cutoff of Aaa to 0, $E(T_0) = 0$, without loss of generality. It can be seen from column 4 to 6, that the average minimum creditworthiness required slightly loosens for Ba and B (and above) as opposed to those

without controlling for the liaison. Furthermore, the differences provide explanation to why we observe significant gaps between NCCP and SCCP in Figure 3.5, especially for B-grade or above categories. We also illustrate this point in the analysis of heterogeneous conditional mean relative thresholds estimates.

3.6.5 Conditional Mean Thresholds

Our main findings are summarized in the Tabel 3.5. We have found a great deal of heterogeneity of overrating biases at both the grade and individual bond level, characterized by the CRA-issuer liaison measure. In the table, we consider the level-specific relative thresholds evaluated at a vector of empirical percentiles of the liaison index. As mentioned earlier, it is often true that threshold biases for Aaa/investment-level are minimal despite a close relationship between firms and CRA because rating of Aaa bonds usually requires strict examination and regulation and meanwhile CRA would suffer severe reputation loss if the default rate of Aaa bonds is higher than industry standard. Therefore, we are permitted to normalize the cutoff for Aaa to be 0 for any R. Now turn to the case 1 in Table 3.5. It can be seen that thresholds for Aa or above, start to loosen at the 70 percentile of the MFOI and dip drastically from 80 percentile. If we could assume the default risk index is uniformly distributed, then bonds with default risk as 2.27 times higher as marginal Aa bonds could have still been rated in the same grade or above. For A grade, conditional mean thresholds begin to shift down around 80 percentile of MFOI at a lesser scale in that the maximal allowable risk is 1.55 times higher than the risk without relaxed thresholds. The heterogeneity is also reflected for when the cutoff starts to loosen up conditional on the liaison. For instance, $\widehat{\Delta}_{0,3}$ begins to decrease as soon as the liaison is at its 20 percentile for Baa grade. For those below Baa, the similar patten can be observed as well. It might be the case that the CRA's criteria for high-yield (riskier) bonds are much easier to be relaxed than for investment-grade bonds. That means that even if the firm and CRA only share somewhat weak relationship, it is still very likely for CRA to overrate bonds of below Baa grade. The maximal allowable default risk, calculated as $\max_R \widehat{\Delta}_{j,0}(R)/\widehat{\Delta}_{j,0}(0)$, ranges from [1.36, 2.27] across grades. Besides, we are also surprised to find out that it is with

the highest probability to be overrated when MFOI is at its 80 percentile, indicating a non-monotonic relationship in liaison. This effect is also captured in Figure 3.6 that depicts the relative thresholds varying with MOFI. From the figure, for blue (Aa) and red (A) lines, they are stable over a large range of small MFOI and then drop drastically around 0.7. As opposed, yellow (Baa) and purple (Ba) dips slightly immediately after MFOI moves away from 0 and also experience the most dramatic drop at 0.7.

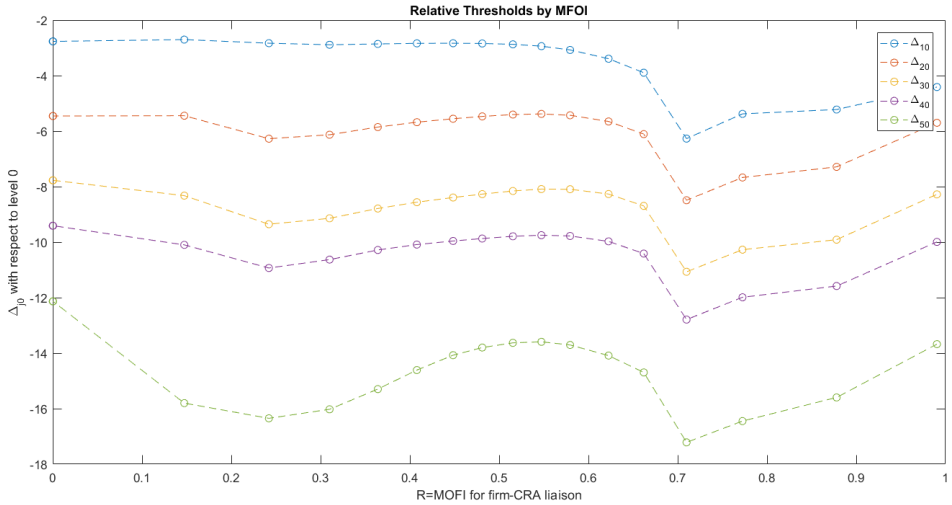
Next, we control for more than MFOI with multiple control variables and the results are presented in the second panel of Table 3.5. Generally speaking, overrating biases could be inferred from the decreasing thresholds estimates as the liaison strengthens. Figure 3.7 plots such relationship. Similar to Figure 3.6, we observed the decreasing effect of the liaison on the conditional cutoff means, implying inflated grades are more likely to be assigned to those who, in some way, are close to the CRA. However, unlike the single control case, the most astonishing feature of the graph is that there are even crossings between mean relative threshold curves. Crossings are counterintuitive because it means that for a given bond, it is easier to be rated at a higher notch than a lower notch given some level of liaison. For instance from the graph, once the liaison index is over 6500, no bonds would be rated below grade A. We admit that this may not be a very precise estimate of the counterfactual since for our sample, there are fewer than 5% observations falling into this range and among which even fewer report grades below A.

Table 3.5: Estimation Results of Relative Thresholds ($\widehat{\Delta}$) at Control Index Percentiles

	$Pctl-.1$	$Pctl-.2$	$Pctl-.3$	$Pctl-.4$	$Pctl-.5$	$Pctl-.6$	$Pctl-.7$	$Pctl-.8$	$Pctl-.9$
					Case 1: the single control				
$\widehat{\Delta}_{0,1}(R)$	-2.740	-2.654	-2.848	-2.780	-2.767	-2.843	-3.331	-6.232	-5.247
$\widehat{\Delta}_{0,2}(R)$	-5.435	-5.396	-6.110	-5.608	-5.378	-5.257	-5.554	-8.421	-7.309
$\widehat{\Delta}_{0,3}(R)$	-7.714	-8.292	-9.113	-8.474	-8.163	-7.949	-8.152	-10.988	-9.933
$\widehat{\Delta}_{0,4}(R)$	-9.331	-10.073	-10.579	-9.990	-9.755	-9.603	-9.847	-12.693	-11.573
$\widehat{\Delta}_{0,5}(R)$	-11.994	-15.709	-15.969	-14.496	-13.614	-13.371	-13.944	-17.122	-15.509
					Case 2: the control index				
$\widehat{\Delta}_{0,1}(\widehat{L})$	-1.908	-2.190	-2.391	-2.459	-2.516	-3.224	-3.617	-3.399	-10.007
$\widehat{\Delta}_{0,2}(\widehat{L})$	-4.691	-4.412	-4.198	-4.296	-4.846	-6.317	-6.956	-6.702	-11.253
$\widehat{\Delta}_{0,3}(\widehat{L})$	-7.077	-7.287	-7.462	-7.746	-8.321	-9.623	-10.097	-10.307	-14.154
$\widehat{\Delta}_{0,4}(\widehat{L})$	-8.751	-8.983	-9.199	-9.533	-10.176	-11.524	-11.931	-10.308	-13.795
$\widehat{\Delta}_{0,5}(\widehat{L})$	-13.516	-13.369	-13.454	-13.736	-14.412	-19.960	-12.028	-10.360	-13.800

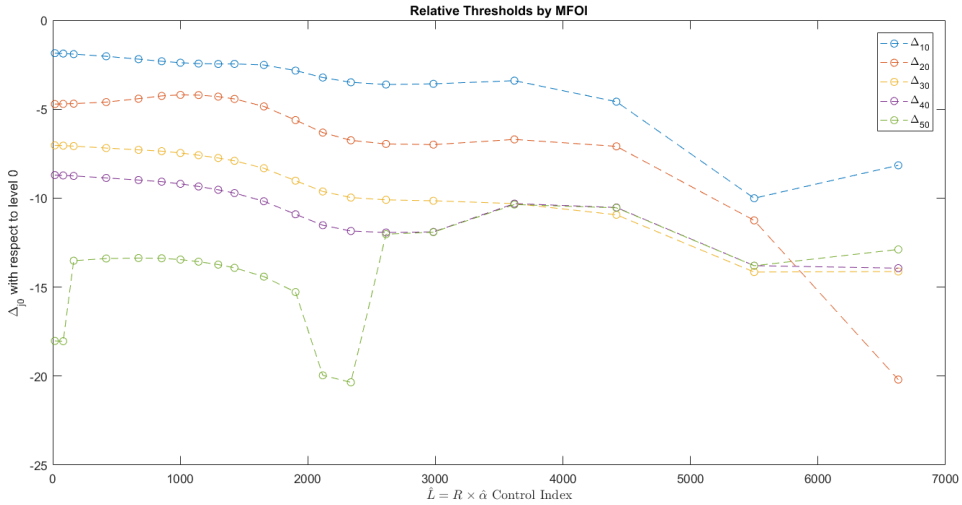
Note: 1. Each $\widehat{\Delta}(\cdot)$ is evaluated at the percentile of R or $\widehat{L}$. Specifically, $pctl-.1$ denotes the 10 percentile. 2. The relative scaled thresholds for notch j is defined as $\Delta_{j,0}(r) = E(T_j - T_0|R = r)$.

Figure 3.6: Heterogeneous Conditional Mean Relative Thresholds—Single Control



Note: 1. Circled points are conditional mean relative thresholds given selected percentile of the control or control index. 2. $\widehat{\Delta}(r)$ is the estimate of $E(T_j - T_0|R = r)$ with the base level being Aaa grade.

Figure 3.7: Heterogeneous Conditional Mean Relative Thresholds—Control Index



Note: 1. Circled points are conditional mean relative thresholds given selected percentile of the control or control index. 2. $\widehat{\Delta}(r)$ is the estimate of $E(T_j - T_0|L = r)$ with the base level being Aaa grade.

3.6.6 Summary of Empirical Results

We begin to compare index estimates from the OProbit and WSLS estimators and confirm that imposing parametric distributional assumptions such as normally distributed thresholds and error terms may significantly downweigh the importance of profitability, asset volatility and bond seniority status, etc., in assessing bond default risks. The conditional

probability rating functions clearly depend on both the default risk index and the liaison index, in an almost monotonic way. We observe that the conditional probability of being rated at a particular grade or above is generally increasing with the liaison measure given some level of default risk. By comparing the average structural probability function with those only conditioning on the default risk, the disparities between NCCP and SCCP provide solid evidence for the endogeneity of bond and firm characteristics motivated in our rating model. Without controlling for the correlation between bond characteristics and stochastic thresholds, one might obtain misleading conditional probabilities and marginal effects, which might be over or under estimated over different ranges of default risk. The distinctive feature of our model is its ability to capture heterogeneous conditional relative thresholds. Our empirical results confirm that there is great amount of unobserved heterogeneity in terms of firm-specific thresholds. In general, as the liaison strengthens by way of increasing common shareholders' investment or others, firms could expect less strict criteria, which in other words, means inflated grading. Furthermore, the overrating biases may be different across grades, with the maximum allowable default risk ratio ranging from [1.36, 2.27]. Amongst all categories, those of A or above experience almost no overrating bias when liaison is moderate; however, significant inflated ratings when liaison exceeds some large number. For bonds of Baa or below, overrating could happen even when liaison is not too strong.

3.7 Conclusions

The credibility of CRAs as third-party information providers for general public investors has been constantly subject to questions especially after the financial crisis 2008. Rating biases of CRAs have been documented in the previous literature. One of the channels that may exert influences on their objectivity is through having common shareholders with bond-issuing firms who are also CRA's clients. In this paper, we consider the bond rating models in the presence of such private information or the firm-CRA liaison. According to our behavioral framework, we note that under this scenario, some of bond characteristics become endogenous as the bond-issuing firms may adjust issuing decisions by forming more accurate beliefs of category thresholds distributions based on the private liaison. Therefore,

we explicitly incorporate the endogenous regressors with heterogeneous thresholds in the empirical model. Through our two-stage semiparametric estimation, the default index parameters and conditional mean thresholds can be identified and estimated.

With a rich bond-level dataset from 2001 to 2008, our empirical evidences support the story of endogenous selection of bond characteristics and the *liaison*-induced omitted variables in rating thresholds. Therefore, controlling for this relationship is necessary to obtain consistent estimators of default index and mean thresholds. More importantly, we find heterogeneous patterns of rating biases across bond grades: those of A or above experience almost no overrating bias when liaison is moderate; however, significant inflated ratings appear after liaison exceeds some threshold. For bonds of Baa or below, overrating could immediately occur as soon as the liaison starts to form.

.1 Identification Proof

Proof of proposition 3.1. We first show the identification of $P_j^s(v)$,

$$\begin{aligned} P_j^s(v) &\equiv \int \{v \leqslant t\} dF_{T_j}(t) \\ &= \int \{v \leqslant t\} dF_{T_j}(t) \\ &= \int \{v \leqslant t\} dF_{T_j|R}(t|r) dF_R(r) \\ &= \int \{v \leqslant t\} dF_{T_j|R}(t|r) dF_R(r) \\ &= \int \{v \leqslant t\} dF_{T_j|R,V}(t|r,v) dF_R(r) \\ &= \int \Pr(Y \leqslant j | V = v, R = r) dF_R(r) \end{aligned}$$

Since $\Pr(Y \leqslant j|V = v, R = r)$ and $F_R(r)$ can be directly estimated from the data, $P_j^s(v)$ is therefore identified. The above argument implicitly use the assumption I.1 and I.2 assuming index structure. Once $P_j^s(\cdot)$ is identified at every point in the support, marginal effect of $P_j^s(\cdot)$ with respect v can be identified as long as the derivative exists. □

Proof of Proposition 3.2 and 3.3. For $\Delta(r)$, the proof of identification resembles that in Klein and Sherman [73]. For each $(x, r) \in \mathcal{X} \times \mathcal{R}$ and $v = x_0 + x'\theta_0$,

$$\begin{aligned} P_j(v, r) &= \Pr(Y \leqslant j | V_0 = v, R = r) \\ &= \Pr(v \leqslant T_j | V_0 = v, R = r) \\ &= \Pr(v - t_j(r) \leqslant U | R = r) \\ &= \Pr(v + \Delta_{j,k}(r) - t_k(r) \leqslant U | R = r) \\ &= \Pr(v + \Delta_{j,k}(r) \leqslant T_k | R = r) \\ &= \Pr(Y \leqslant k | V = v + \Delta_{j,k}(r), R = r) \\ &= P_k(v + \Delta_{j,k}(r), r) \end{aligned}$$

For point identification of $\Delta_{j,k}(r)$, note that $P_k(, r)$ is invertible because it is the conditional distribution function of U given $R = r$. So $\Delta_{j,k}(r) = P_k^{-1}(P_j(v, r), r) - v$, where

$P_k^{-1}(P_k(v,r),r) = v$. Then by construction, $\Delta_{j,k}(r)$ is identified. □

.2 Proof of Asymptotic Theorems

This section first states the asymptotic assumptions in Appendix .2.1 and then gives formal proofs of Theorem 3.2 in Appendix .2.2. For Theorem 3.1, see Ichimura and Lee [58], Ichimura [57]. All supporting lemmas are given in Appendix .2.3.

.2.1 Asymptotic Assumptions

A-A.1. DGP. $\{(Y_i, X_i', R_i', \boldsymbol{T}_i')\}_{i=1}^N \in (\mathcal{Y}, R^{dx}, R, R^J)$ is an i.i.d. vector of random variables defined on a complete probability space $(\Omega, \mathcal{F}, P)$, where (Y_i, X_i', R_i') are observed and $\boldsymbol{T}_i'$ are unobserved.

A-A.2. Smoothness. For each $j \in \mathcal{Y}$ and $(v,r) \in \mathbb{R} \times \mathcal{R}$, $0 < P_j(v,r) < 1$. The CDF F_V and F_R has the uniformly continuous and bounded Radon-Nikodym second order density derivatives with respect to Lebesgue measure. i). f_V is continuous in v and $f_{V|R}$ is continuous in (v,r). ii). There exists $C > 0$ such that $\inf_{\mathbb{R}_0} f_V > C$ and $\inf_{\mathbb{R}_0} f_{V|R} > C$.

A-A.3. Dominance. For each $j \in \mathcal{Y}$ any $r \in \mathcal{R}$, $P_j(\cdot, r)$ has all partial derivatives up to 3rd order. Let $\nabla^l P_j(v,r) \equiv \frac{\partial^l P_j(v,r)}{(\partial v)^l}$ where $l = 1,2,3$. $\nabla^l P_j(\cdot, r)$ is uniformly bounded and Lipschitz continuous on $\mathbb{R}$: for all $v, \widetilde{v} \in \mathbb{R}$, $|\nabla^l P_j(v,r) - \nabla^l P_j(\widetilde{v}, r)| \leqslant C||v - \widetilde{v}||$, for some constant $C > 0$, where $||\cdot||$ is the Euclidean norm.

A-A.4. Kernel. For some integer ν, the univariate symmetric kernel function $k : \mathbb{R} \to (0,1)$, satisfies i). $\int u^i k(u) du = \delta_{i0}$, for $i = 0, 1, \cdots, \nu - 1$, where δ_{ij} is the Kronecker's delta. ii). $\int u^\nu k(u) du < \infty$. iii). $k(u) = O((1 + u^{1+u+\varepsilon})^{-1})$ for some $\varepsilon > 0$.

A-A.5. Bandwidth. As $N \to \infty$, then $h_i \to 0$, $Nh_i^4 \to \infty$, for $i = 1, 2$, $\sqrt{N}h_1^6 \to 0$ and $\sqrt{Nh_2}h_2^4 \to 0$.

A-A.6. Parameter space. i). $\theta_0 \in \Theta_0 \subset \mathbb{R}^d$, where Θ_0 is the interior of the compact support Θ. ii). for each $r \in \mathcal{R}_0$, $\Delta(r) \in \mathcal{D}(r) \subset \mathbb{R}^{d_J - 1}$, where $\mathcal{D}(r)$ is the interior of the compact support of $\Delta(r)$.

A-A.1 reiterates the data generating process, already embodied in our model (3.3). We do not need X and R to be compactly supported as the trimming indicator will guarantee the density denominators away from 0. A-A.2 and A-A.3 are regularity conditions usually appearing in nonparametric estimators. They indicate that densities and conditional expectations are smooth enough and have partial derivatives up to 3rd order with respect to the index V. A-A.4 is standard in kernel estimation. In this paper, the second-order kernels, $\nu = 2$, mostly suffices to reduce the asymptotic bias. A-A.5 concerns bandwidths and window parameters. Silverman's rule of thumb bandwidth, e.g. $h_i = 1.06 \times std \times N^{-r_i}$, for $i = 1, 2$, is being used. A-A.6 restricts support of the finite and infinite-dimensional parameters to be compact given point identification.

.2.2 Proofs

Before proving the second stage formally, we begin by introducing some notational abbreviations. Since we are localizing around r, then for each r in a compact subset of the support of R, we suppress the dependency on r for simplicity but readers should be advised that almost all objects are functions of this control variable. Then we simplify the expression for the sample gradients evaluating at the truth to be Eq. (30),

$$\widehat{G}_N(\widehat{V}) \equiv N^{-1} \sum_{i=1}^{N} \widehat{\tau}_i' \widehat{\Psi}(\widehat{V}_i, \Delta_0) \tag{30}$$

where, admitted of a slight abuse of notation, we redefine the trimming function as $\mathbf{1} \otimes \tau_i(r)$ to incorporate multi-dimensional feature of the restrictions, where $\mathbf{1}$ denotes a $J-1$-dimensional vector of ones and $\otimes$ is the Kronecker product. And the estimated summand is give by $\widehat{\Psi}(\widehat{V}_i, \Delta_0) = (\widehat{\psi}_0, \widehat{\psi}_1, \cdots, \widehat{\psi}_{J-2})'$ and for each j, we have

$$\widehat{\psi}_j(\widehat{V}, \Delta_0) = [\widehat{P}_j(\widehat{V}_i) - \widehat{P}_{j+1}(\widehat{V}_i + \Delta_0)] \widehat{P}_{j+1}'(\widehat{V}_i + \Delta_0), \; j = 0, 1 \cdots, J-2 \tag{31}$$

Likewise, we redefine the limiting gradient and its components by letting $\Psi(v, \Delta) = (\psi_0, \psi_1, \cdots, \psi_{J-2})'$,

$$\psi_j = \psi_j(v, \Delta) = [P_j(v) - P_{j+1}(v + \Delta)] P_{j+1}'(v + \Delta), \; j = 0, 1 \cdots, J-2$$

Then the limiting gradient evaluating at the truth is defined in Eq. (32),

$$G_N = \frac{1}{N}\sum_{i=1}^{N} \tau_i' \Psi(V_i, \Delta_0) \tag{32}$$

Next, define the Hessian matrix $\widehat{H}_N(\widehat{V}_i) = \partial \widehat{G}_N(\widehat{V}_i)/\partial \Delta'$, stated in Eq. (33), denoting the derivative of the gradients with respect to Δ.

$$\widehat{H}_N(\widehat{V}) \equiv N^{-1} \sum_{i=1}^{N} \widehat{\tau}_i' \widehat{h}(\widehat{V}_i, \Delta_0) \tag{33}$$

where P_j' and P_j'' indicate the first and second derivatives of $P_j(\cdot)$.

$$\widehat{h}_j(\widehat{V}, \Delta_0) = [\widehat{P}_j(\widehat{V}_i) - \widehat{P}_{j+1}(\widehat{V}_i + \Delta_0)]\widehat{P}_{j+1}'(\widehat{V}_i + \Delta_0) - \widehat{P}_{j+1}'(\widehat{V}_i + \Delta_0)\widehat{P}_{j+1}'(\widehat{V}_i + \Delta_0)' \tag{34}$$

The limiting Hessian is analogously defined as H_N for brevity.

Theorem 3.2 considers the consistency and asymptotic normality of localized relative thresholds given $R = r$ in a compact support $\mathcal{R}_0$. For simplicity, we suppress the dependency on r for demonstrative purpose. A-A.6 implies that the gradient vector to the minimization problem in (3.13) is set to 0 when evaluated at $\widehat{\Delta}$. In Eq. (35), by Taylor expansion around Δ_0, we decompose the gradients, $\widehat{G}_N(\widehat{\Delta})$, into two components.

$$0 = \widehat{G}_N(\widehat{\Delta}) = \widehat{G}_N(\Delta_0) + \widehat{H}_N(\Delta^+)(\widehat{\Delta} - \Delta_0) \tag{35}$$

where $\widehat{H}_N(\cdot)$ denotes the derivative of the gradients with respect to Δ and Δ^+ is the mean value in between $\widehat{\Delta}$ and Δ_0. Then we can primarily work with $(\widehat{\Delta} - \Delta_0)$ thereafter.

$$(\widehat{\Delta} - \Delta_0) = \widehat{H}_N(\Delta^+)^{-1}\widehat{G}_N$$

Now, we first show that the estimated Hessian, $\widehat{H}(\Delta^+)$, converging in probability to the true one, uniform in Δ by Lemma .1, i.e. $\widehat{H}_N(\Delta^+) \xrightarrow{\text{P}} H_0$ where $H_0 \equiv E[H_N(\Delta_0)]$. Next, we prove that the estimated gradient evaluated at the truth converging in distribution to a multivariate normal at the nonparametric rate of $\sqrt{Nh}$. Since the gradients consist

of estimated trimming functions, semiparametric conditional probabilities as well as the estimated index, we decompose $\widehat{G}_N$ into multiple terms as below and relate it to the true gradients. To begin with,

$$\widehat{G}_N(\widehat{V}, \Delta_0) \quad = \quad I_1 + I_2 + I_3 + U_1 + U_2 + U_3$$

where in particular,

$$\begin{aligned}
I_1 &= N^{-1} \sum_{i=1}^{N} \tau_i' \Psi(V_i, \Delta_0) \\
I_2 &= N^{-1} \sum_{i=1}^{N} \tau_i' [\widehat{\Psi}(\widehat{V}_i, \Delta_0) - \widehat{\Psi}(V_i, \Delta_0)] \\
I_3 &= N^{-1} \sum_{i=1}^{N} \tau_i' [\widehat{\Psi}(V_i, \Delta_0) - \Psi(V_i, \Delta_0)] \\
U_1 &= N^{-1} \sum_{i=1}^{N} (\widehat{\tau}_i - \tau_i)' \Psi(V_i, \Delta_0) \\
U_2 &= N^{-1} \sum_{i=1}^{N} (\widehat{\tau}_i - \tau_i)' [\widehat{\Psi}(\widehat{V}_i, \Delta_0) - \widehat{\Psi}(V_i, \Delta_0)] \\
U_3 &= N^{-1} \sum_{i=1}^{N} (\widehat{\tau}_i - \tau_i)' [\widehat{\Psi}(V_i, \Delta_0) - \Psi(V_i, \Delta_0)]
\end{aligned}$$

The first three terms, $I_1 - I_3$, reflect the components eliminating the estimation variability from the trimming. As opposed, the last three terms, $U_1 - U_3$, consist those arising from the estimated index. For example, I_1 is true gradient with true trimming functions and is equal to zero as $\psi_j(V_i, \Delta_0) = 0$ for each j. Likewise, $U_1 = 0$. I_2 concerns the pass-through of estimation variability due to the unknown index in the first stage and this effect is compounded with that of estimating the nonparametric conditional probability functions and their derivatives. Fortunately, I_2 converges at a faster parametric rate and vanishes in the limit, as proved by Lemma .2. I_3 reflects the variability arising from the nonparametric conditional expectation functions and derivative estimation, contributing to the limiting variance, as shown in Lemma .3. Both U_2 and U_3 are converging to 0 at a rate faster than $\sqrt{N}$, let alone the nonparametric rate here, therefore vanishing in the limit, see Lemma .4. Hence, the standard central limit theorem (CLT) would apply with an adjustment on the

rate.

Lemma .1 (Uniform convergence of $\widehat{H}_N$)**.**

$$\sup_{\Delta \in \mathcal{D}(r)} |\widehat{H}_N(\widehat{V}, \Delta, r) - E[H_N(V, \Delta, r)]| = o_p(1)$$

also note that $H_0(r) = E[H_N(V, \Delta_0, r)]$.

Proof. Given $R = r$ and the corresponding compact support $\mathcal{D}(r)$ and we omit the dependency on r and Δ,

$$\sup_{\Delta \in \mathcal{D}(r)} |\widehat{H}_N(\widehat{V}) - H_0| \leqslant \underbrace{\sup_{\Delta \in \mathcal{D}(r)} |\widehat{H}_N(\widehat{V}) - \widehat{H}_N(V)|}_{S_1} + \underbrace{\sup_{\Delta \in \mathcal{D}(r)} |\widehat{H}_N(V) - H_N(V)|}_{S_2} + \underbrace{\sup_{\Delta \in \mathcal{D}(r)} |H_N(V) - EH_N|}_{S_3}$$

We need show that all three terms are converging to 0. $S_1 = o_p(1)$ is given by applying Taylor expansion around v in conjunction with the convergence of V in Lemma 8. For S_2, since the estimated Hessian consists of only $\bigtriangledown^d \widehat{P}_j(\widehat{V})$ or $\bigtriangledown^d \widehat{P}_{j+1}(\widehat{V} + \Delta)$ for $d = 0, 1, 2$ and $j = 0, 1, \cdots J - 2$, the uniform convergence is guaranteed by Lemma A 4. $S_3 = o_p(1)$ is given by standard argument of LLN.

□

Lemma .2 (I_2)**.**

$$\sqrt{Nh} I_2 = o_p(1)$$

Proof.

$$I_2 \quad = \quad N^{-1} \sum_{i=1}^{N} \tau_i' [\widehat{\Psi}(\widehat{V}_i, \Delta_0) - \widehat{\Psi}(V_i, \Delta_0)]$$

Note that $\widehat{\Psi}(v)$ is continuously differential in v since the kernel is smooth guaranteed by A-A-4. By mean value theorem,

$$\widehat{\Psi}(\widehat{V}_i) - \widehat{\Psi}(V_i) = \widehat{\Psi}'(V_i^+)(\widehat{V}_i - V_i), \quad V_i^+ \in (V_i, \widehat{V}_i)$$

By A-A.2 and A-A.3, $\widehat{\Psi}'(V_i^+)$ is uniformly bounded in v on a compact set. By Lemma A 8, it can be seen that $\sqrt{N^{-1}\sum_{i=1}^N(\widehat{V}_i - V_i)^2} = O(N^{-1/2})$. By Cauchy-Schwartz inequality,

$$\sqrt{Nh}I_2 \leqslant \sqrt{Nh}C\sqrt{N^{-1}\sum_{i=1}^N(\widehat{V}_i - V_i)^2} = O(h)$$

Therefore, I_2 is vanishing faster as opposed to the nonparametric rate. □

Lemma .3 (I_3)**.**

$$I_3 \quad = \quad N^{-1}\sum_{i=1}^N \boldsymbol{\tau}_i'\xi_i + o_p\left(\frac{1}{\sqrt{Nh}}\right)$$

where ξ_i is defined in Eq. (3.23)*.*

Proof.

$$I_3 \quad = \quad N^{-1}\sum_{i=1}^N \boldsymbol{\tau}_i'[\widehat{\Psi}(V_i, \Delta_0) - \Psi(V_i, \Delta_0)]$$

For exposition, we take $j = 0$ as an example; for $j = 1, \cdots, J-2$, the same calculation applies over.

$$\widehat{\psi}_{0,i} = P_{1,i}'[(\widehat{P}_0(V_i, r) - P_0(V_i, r)] - P_{1,i}'[\widehat{P}_1(V_i + \Delta, r) - P_0(V_i, r)]$$

By Lemma A 5, we can instead work with $\widetilde{\psi}_i$, defined in Eq. (36), because $\sqrt{Nh}|\widehat{\psi}_{0,i} - \widetilde{\psi}_{0,i}| = o_p(1)$

$$\widetilde{\psi}_i \quad = \quad \widetilde{\psi}_i^1 + \widetilde{\psi}_i^2 \tag{36}$$

$$\widetilde{\psi}_i^1 \quad = \quad g(V_i, r)^{-1}P_{1,i}'(N-1)^{-1}\sum_{j\neq i}^{N-1} K(R_j - r)K(V_j - V_i)\left[Y_{0,j} - P_0(V_i, r)\right] \tag{37}$$

$$\widetilde{\psi}_i^2 \quad = \quad g(V_i + \Delta, r)^{-1}P_{1,i}'(N-1)^{-1}\sum_{j\neq i}^{N-1} K(R_j - r)K(V_j - V_i - \Delta_0)\left[Y_{1,j} - P_0(V_i, r)\right] \tag{38}$$

where $Y_{k,i} = \{Y_i \leqslant k\}$, for $k = 0, 1$. Next we show that $N^{-1}\sum_i^N \widetilde{\psi}_i^k, k = 1, 2$ can be

represented in the form of second-order U-statistics.

$$\begin{aligned} N^{-1}\sum_{i=1}^{N}\widetilde{\psi}_i^1 &= \frac{1}{N(N-1)}\sum_{j\neq i}\tau_i g(V_i,r)^{-1}P'_{1,i}K(R_j-r)K(V_j-V_i)\left[Y_{0,j}-P_0(V_i,r)\right] \\ &= \binom{N}{2}^{-1}\sum_{j>i}(\phi_{ij}+\phi_{ji})/2 \end{aligned}$$

where

$$\begin{aligned} \phi_{ij} &= \tau_i g(V_i,r)^{-1}P'_{1,i}K(R_j-r)K(V_j-V_i)\left[Y_{0,j}-P_0(V_i,r)\right] \\ \phi_{ji} &= \tau_j g(V_j,r)^{-1}P'_{1,j}K(R_i-r)K(V_i-V_j)\left[Y_{0,i}-P_0(V_j,r)\right] \end{aligned}$$

It is obvious that $E|\phi_{ij}^2+\phi_{ji}^2| = o(N)$ by A-A.

$$\begin{aligned} E(\phi_{ij}|i) &= O(h^2) \\ E(\phi_{ji}|i) &= \tau_i g(V_i,r)^{-1}P'_{1,i}K(R_i-r)f(V_i)[Y_{0,i}-P(V_i,r)]+O(h^2) \end{aligned}$$

As long as $r > 1/5$, the bias of order h^2 would vanish in the limit. By the $U-$statistic projection theory, one can show that

$$N^{-1}\sum_{i=1}^{N}\widetilde{\psi}_i^1 = N^{-1}\sum_{i=1}^{N}\tau_i g(V_i,r)^{-1}P'_{1,i}K(R_i-r)f(V_i)[Y_{0i}-P(V_i,r)]+o_p\left(\frac{1}{\sqrt{Nh}}\right) \quad (39)$$

In the same manner, it is easy to obtain an analogous expression for $N^{-1}\sum_{i=1}^{N}\widehat{\psi}_i^2$ like Eq. (40) where every V_i is replaced with the shifted index $V_i+\Delta_0$ and Y_{0i} with Y_{1i}.

$$N^{-1}\sum_{i=1}^{N}\widetilde{\psi}_i^2 = N^{-1}\sum_{i=1}^{N}\tau_i g(V_i+\Delta_0,r)^{-1}P'_{1,i}K(R_i-r)f(V_i+\Delta_0)[Y_{1i}-P(V_i,r)]+o_p\left(\frac{1}{\sqrt{Nh}}\right) \quad (40)$$

To sum up, I_3 can be represented as the sum of two $U-$statistic projections plus some remainders converging faster to 0 than the nonparametric rate.

$$N^{-1}\sum_{i=1}^{N}\widehat{\psi}_{0,i} = N^{-1}\sum_{i=1}^{N}\xi_{0,i}+o_p\left(\frac{1}{\sqrt{Nh}}\right)$$

where

$$\xi_{0,i} = \tau_i P'_{1,i} K(R_i - r) \left\{ \frac{f(V_i)}{g(V_i, r)}[Y_{0,i} - P(V_i, r)] - \frac{f(V_i + \Delta_0)}{g(V_i + \Delta_0, r)}[Y_{1,i} - P(V_i, r)] \right\}$$

The above argument would straightforwardly apply to each $j = 1, \cdots, J-2$ and then we collect them in a vector $\xi_i = (\xi_{0,i}, \xi_{1,i}, \cdots, \xi_{J-2,i})'$. □

Lemma .4 (U_2 & U_3)**.**

$$\sqrt{Nh}U_2 = o_p(1), \quad \sqrt{Nh}U_3 = o_p(1)$$

Proof. Recall that

$$\begin{aligned} U_2 &= N^{-1} \sum_{i=1}^{N} (\widehat{\tau}_i - \tau_i)' [\widehat{\Psi}(\widehat{V}_i, \Delta_0) - \widehat{\Psi}(V_i, \Delta_0)] \\ U_3 &= N^{-1} \sum_{i=1}^{N} (\widehat{\tau}_i - \tau_i)' [\widehat{\Psi}(V_i, \Delta_0) - \Psi(V_i, \Delta_0)] \end{aligned}$$

For U_2, similar to Lemma .2, under A-A.2 and A-A.3, it can be shown that

$$\sqrt{Nh}I_2 \leqslant \sqrt{Nh} \sqrt{N^{-1} \sum_{i=1}^{N} (\widehat{\tau}_i - \tau_i)^2} \sqrt{N^{-1} \sum_{i=1}^{N} (\widehat{V}_i - V_i)^2} \leqslant C\sqrt{Nh} \sqrt{N^{-1} \sum_{i=1}^{N} (\widehat{V}_i - V_i)^2} = o_p(1)$$

where $C = \max_i |\widehat{\tau}_i - \tau_i|$ a upper bound of constant.

For U_3, it is implied by Lemma .3 that

$$\sqrt{Nh}U_3 = \sqrt{Nh}I_3 \max_i |\widehat{\tau}_i - \tau_i| = o_p(1)$$

Since from Lemma A 7, it turns out that $\max_i |\widehat{\tau}_i - \tau_i| = O(N^{-1/2+\epsilon})$[CONFIRM], for some $\epsilon > 0$. □

.2.3 Intermediate Lemmas

Notation Let $\widehat{f}(v, \theta) \equiv N^{-1} \sum_{i=1}^{N} K_h(V_i(\theta) - v) Y_i$ and $\widehat{g}(v, \theta) \equiv N^{-1} \sum_{i=1}^{N} K_h(V_i(\theta) - v)$.

Lemma 4 (Convergence rates)**.** *For V a d-dimensional vector of continuous random*

variables with density g_V. *Let* $\nabla_\theta^l g_V$ *be the* l^{th} *partial derivatives of* g_V *with respect to* θ, *and* $\nabla_\theta^0 g_V = g_V$. *Let* $\widehat{g}_V$ *represents the estimator of* g_V. *Then, for* θ *in a compact set and* v *in a compact subset of the support of* V, *the following rates hold for* $l = 0, 1, 2$,

$$i). \qquad \sup_{v,\theta} E\left\{\left[\nabla_\theta^d \widehat{g}_V(v,\theta) - E\left(\nabla_\theta^d \widehat{g}_V(v,\theta)\right)\right]^2\right\} = O(\frac{1}{Nh^{2d+2l+1}})$$

$$ii). \qquad \sup_{v,\theta}\left|E\left(\nabla_\theta^d \widehat{g}_V(v,\theta) - \nabla_\theta^d g_V(v,\theta)\right)\right| = O(h^2)$$

The proof follows from Lemma 3 in Klein and Shen [69] where they consider the univariate case for $d = 1$.

Lemma 5 (Double convergence)**.** *Suppose* θ *in a compact set and* v *in a compact subset of the support of a* $d-$*dimensional vector of continuous variables* V, *if* $1/8 < r < 2/d$, *then*

$$\sqrt{N}\left|\widehat{E}(Y|v,\theta) - E(Y|v,\theta)\right| = \sqrt{N}\left|\widehat{f}(v,\theta) - E(Y|v,\theta)\widehat{g}(v,\theta)\right|/g(v,\theta) + o_p(1)$$

Lemma 6 (Bahadur Representation from Bahadur [9])**.** *Suppose that* $\widehat{q}_V(\lambda)$ *and* $q_V(\lambda)$ *are estimated and true quantile functions of a* $d-$*dimensional continuous vector of random variable* V *evaluated at a vector of* $\lambda \in [0,1]^d$.

$$\sqrt{N}(\widehat{q}_V(\lambda) - q_V(\lambda)) = N^{-1}\sum_{i=1}^{N} B_i + o_p(1)$$

where $B_i = (B_{1i}, B_{2i}, \cdots, B_{di})'$ *and for each* $j = 1, 2, \cdots, d$,

$$B_{ji} = \frac{\mathbf{1}[V_{ji} \leqslant q_{V_j}(\lambda_j)] - \lambda_j}{g_{V_j}(v)} \tag{41}$$

Lemma 7 (Estimated trimming from Lemma 3 in Klein and Shen [70])**.** *Suppose that* W_i *is a random variable or function satisfying that* $m(q) =\equiv E[\tau_i(q)W_i]$ *is bounded and continuously differentiable in* q *where* q *denotes the vector of true quantiles of the continuous variable* V. *Then*

$$N^{-1/2}\sum_{i=1}^{N}[\tau(\widehat{q}) - \tau_i(q)]W_i = m'(q)N^{-1/2}\sum_{i=1}^{N} B_i + o_p(1)$$

where B_i is defined in (41) and $m'(\cdot)$ is the derivative of m.

Lemma 8 (Index)**.** *For each i, $m(\cdot)$ is a bounded and continuously differential function.*

$$i). \qquad |\widehat{V}_i - V_i| = o_p(N^{-1/2})$$
$$ii). \qquad |m(\widehat{V}_i) - m(V_i)| = o_p(N^{-1/2})$$

Proof. For i). $\widehat{V}_i = X_{0i} + X'\widehat{\theta} = V_i + X_i'(\widehat{\theta} - \theta_0)$, so $\sqrt{N}(\widehat{V}_i - V_i) = X_i'\sqrt{N}(\widehat{\theta} - \theta) = o_p(1)$ as indicated from Theorem 3.1. For nonlinear indices, such as $V(\widetilde{X}, \theta)$, if $V(\cdot)$ is twice continuously differentiable in θ, then i) also holds. For ii). It is implied by the Taylor expansion around V_i that $m(\widehat{V}_i) - m(V_i) = m'(V_i)(\widehat{V}_i - V_i) + o_p(N^{-1})$. According to i), ii) holds automatically. □

.3 Grid Search of Initial Values of $\Delta(\cdot)$

Grid Search of Klein and Sherman [73]. After obtaining a consistent estimator of the index, $\widehat{V}_i = X_{0i} + \widetilde{X}_i'\widehat{\theta}$ (and $\widehat{L} = R_{0i} + \widetilde{R}_i'\widehat{\alpha}$), we then estimate the relative differences of thresholds for each value $r \in \mathcal{R}$ or $l \in \mathcal{L}$ by MDE, implied by conditional shift restrictions in Proposition 3.2. As MDE is very sensitive to the choices of starting values, it would be useful to obtain a preliminary estimator which is very close to the true values. To this end, we suggest to experiment with a fast and easy-to-implement grid search estimator to be the initial values upon which the MDE can be subsequently programmed.

As inspired by Klein and Sherman [73], they document a grid search estimator for the constant thresholds in the presence of only exogenous covariates. They claim the grid search method can be easily implemented and fast computed without optimization. Likewise, we extend this approach to our model by first localizing around $R_i = r$ and then searching for $\Delta(r)$. The essential idea is that given $R_i = r$, according to the conditional shift restrictions, $\Delta_{j,j+1}(r)$ can be estimated $\widetilde{v} - v$ where $\widetilde{v}$ and v satisfy $P_j(v, r) = P_{j+1}(\widetilde{v}, r)$. For estimation, we average all possible differences $\widehat{V}_i - \widetilde{V}_i$ for which $\widehat{P}_j(\widehat{V}_i, r)$ and $\widehat{P}_{j+1}(\widetilde{V}_i, r)$ are sufficiently close to each other. In Klein and Sherman's terminology, we search over the overlap of the target set and the grid of points defined respectively.

We first define the overlapping range given $R_i = r$,

$$\begin{aligned} L(r) &:= \max[\min_i(\widehat{P}_j(\widehat{V}_i, r), \min_i(\widehat{P}_{j+1}(\widehat{V}_i, r)] \\ H(r) &:= \min[\max_i(\widehat{P}_j(\widehat{V}_i, r), \max_i(\widehat{P}_{j+1}(\widehat{V}_i, r)] \end{aligned}$$

Define the local target set for level j,

$$\mathcal{S}_T(r) = \left\{\widehat{V}_i : \widehat{P}_L \leqslant \widehat{P}_j(\widehat{V}_i, r) \leqslant \widehat{P}_H\right\}$$

where $\widehat{P}_L$ and $\widehat{P}_H$ denote the corresponding α and $(1-\alpha)$ quantiles of estimated probabilities $\widehat{P}_j, \widehat{P}_{j+1}$that fall in the range $[L(r), H(r)]$. Define $\mathcal{S}_R = \left\{\widehat{V}_i : \widehat{P}_L \leqslant \widehat{P}_{j+1}(\widehat{V}_i, r) \leqslant \widehat{P}_H\right\}$.

For any $p > 1/2$, the grid consists points for which the distance between adjacent ones is $o_p(N^{-1/2})$,

$$\mathcal{S}_G(r) = \left\{\widehat{V}_L(r) + [\widehat{V}_H(r) - \widehat{V}_L(r)]k/N^p, k = 1, 2, \cdots, N^p\right\}$$

where $\widehat{V}_L(r)$ denotes the largest estimated index value smaller than the smallest estimated index value in $\mathcal{S}_R$ and $\widehat{V}_H(r)$ denotes the smallest estimated index value larger than the largest estimated index value in $\mathcal{S}_R$.

In our model, since we are searching every r in the compact support, it is possible that no overlaps between the target set and grid exist in a particular sample or the common points are too few to generate reasonable variances for some set in $\mathcal{R}$. So only the grid search method alone might not produce the desired estimates of relative thresholds. To counteract this shortcoming, we suggest to perform the MDE with initial values given by the grid search.

Bibliography

[1] Daniel Ackerberg, C Lanier Benkard, Steven Berry, and Ariel Pakes. Econometric tools for analyzing market outcomes. *Handbook of econometrics*, 6:4171–4276, 2007.

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

[3] Anat R Admati and Paul Pfleiderer. The wall street walk and shareholder activism: Exit as a form of voice. *Review of Financial Studies*, page hhp037, 2009.

[4] Hyungtaik Ahn, Hidehiko Ichimura, and James L Powell. Simple estimators for monotone index models. 1996.

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

[6] Christopher Alessi, Roya Wolverson, and Mohammed Aly Sergie. The credit rating controversy. *Council on Foreign Relations*, 2013.

[7] Joseph G Altonji and Rosa L Matzkin. Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, pages 1053–1102, 2005.

[8] Manuel Arellano and Stephen Bond. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297, 1991.

[9] R Raj Bahadur. A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37(3):577–580, 1966.

[10] Martin Neil Baily and Barry P Bosworth. Us manufacturing: Understanding its past and its potential future. *The Journal of Economic Perspectives*, 28(1):3–25, 2014.

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

[12] 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.

[13] Marshall E Blume, Felix Lim, and A Craig MacKinlay. The declining credit quality of us corporate debt: Myth or reality? *The journal of finance*, 53(4):1389–1413, 1998.

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

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

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

[17] John C Bogle. *The battle for the soul of capitalism*. Yale University Press, 2005.

[18] Bryan W Brown. The identification problem in systems nonlinear in the variables. *Econometrica: Journal of the Econometric Society*, pages 175–196, 1983.

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

[20] John Y Campbell and Glen B Taksler. Equity volatility and corporate bond yields. *The Journal of Finance*, 58(6):2321–2350, 2003.

[21] Richard Cantor, Frank Packer, et al. The credit rating industry. *Quarterly Review*, (Sum):1–26, 1994.

[22] 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.

[23] Songnian Chen. Semiparametric estimation of a location parameter in the binary choice model. *Econometric Theory*, 15(01):79–98, 1999.

[24] Songnian Chen. Rank estimation of a location parameter in the binary choice model. *Journal of Econometrics*, 98(2):317–334, 2000.

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

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

[27] Jess Cornaggia and Kimberly J Cornaggia. Estimating the costs of issuer-paid credit ratings. *Review of Financial Studies*, 26(9):2229–2269, 2013.

[28] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

[29] Mitali Das. Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal of Econometrics*, 124(2):335–361, 2005.

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

[31] Jan De Loecker. Recovering markups from production data. *International Journal of Industrial Organization*, 29(3):350–355, 2011.

[32] Xavier DHaultfœuille and Philippe Février. Identification of nonseparable triangular models with discrete instruments. *Econometrica, Forthcoming*, 2015.

[33] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

[34] 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.

[35] 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.

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

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

[38] Zhutong Gu. Identification and testing of nonparametric production functions without hicks-neutral productivity shocks. Working Paper, 2017.

[39] Zhutong Gu. Additive separability and unobserved excess heterogeneity: A nonparametric test with average structural functions. Working paper, 2017.

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

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

[42] Robert E Hall. The relation between price and marginal cost in us industry. *Journal of Political Economy*, 96(5):921–947, 1988.

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

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

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

[46] Stefan Hoderlein and Enno Mammen. Identification of marginal effects in nonseparable models without monotonicity. *Econometrica*, 75(5):1513–1518, 2007.

[47] 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.

[48] Stefan Hoderlein and Robert Sherman. Identification and estimation in a correlated random coefficients binary response model. *Journal of Econometrics*, 2015.

[49] Stefan Hoderlein, Liangjun Su, and Halbert White. Specification testing for nonparametric structural models with monotonicity in unobservables. *V UCSD Department of Economics Working Paper*, 2011.

[50] Stefan Hoderlein, Liangjun Su, Halbert L White Jr, and Thomas Tao Yang. Testing for monotonicity in unobservables under unconfoundedness. *Available at SSRN 2448681*, 2014.

[51] Joel L Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica: journal of the Econometric Society*, pages 505–531, 1992.

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

[53] James O Horrigan. The determination of long-term credit standing with financial ratios. *Journal of Accounting Research*, pages 44–62, 1966.

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

[55] Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282, 2010.

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

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

[58] 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.

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

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

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

[62] John Xuefeng Jiang, Mary Harris Stanford, and Yuan Xie. Does it matter who pays for bond ratings? historical evidence. *Journal of Financial Economics*, 105(3):607–621, 2012.

[63] Robert S Kaplan and Gabriel Urwitz. Statistical models of bond ratings: A methodological inquiry. *Journal of Business*, pages 231–261, 1979.

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

[65] Maximilian Kasy. Identification in triangular systems using control functions. *Econometric Theory*, 27(03):663–671, 2011.

[66] Simi Kedia, Shivaram Rajgopal, and Xing Zhou. Does it matter who owns moody's, 2016.

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

[68] Lawrence Robert Klein. textbook of econometrics. 1953.

[69] Roger Klein and Chan Shen. Bias corrections in testing and estimating semiparametric, single index models. *Econometric Theory*, 26(06):1683–1718, 2010.

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

[71] Roger Klein and Francis Vella. A semiparametric model for binary response and continuous outcomes under index heteroscedasticity. *Journal of Applied Econometrics*, 24(5):735–762, 2009.

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

[73] Roger W Klein and Robert P Sherman. Shift restrictions and semiparametric estimation in ordered response models. *Econometrica*, 70(2):663–691, 2002.

[74] Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

[75] Pepa Kraft. Rating agency adjustments to gaap financial statements and their effect on ratings and credit spreads. *The Accounting Review*, 90(2):641–674, 2014.

[76] Pepa Kraft. Do rating agencies cater? evidence from rating-based contracts. *Journal of Accounting and Economics*, 59(2):264–283, 2015.

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

[78] Arthur Lewbel. Semiparametric estimation of location and other discrete choice moments. *Econometric Theory*, 13(01):32–51, 1997.

[79] Arthur Lewbel. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97(1):145–177, 2000.

[80] Arthur Lewbel. Ordered response threshold estimation. *Unpublished working paper, Boston College*, 2003.

[81] 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.

[82] Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

[83] Oliver Linton and Jens Perch Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, pages 93–100, 1995.

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

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

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

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

[88] Charles F Manski. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, 27(3):313–333, 1985.

[89] Jacob Marschak and William H Andrews. Random simultaneous equations and the theory of production. *Econometrica, Journal of the Econometric Society*, pages 143–205, 1944.

[90] 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.

[91] Rosa L Matzkin. Nonparametric estimation of nonadditive random functions. *Econometrica*, pages 1339–1375, 2003.

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

[93] Rosa L Matzkin. Identification in nonparametric simultaneous equations models. *Econometrica*, pages 945–978, 2008.

[94] Jürgen Maurer, Roger Klein, and Francis Vella. Subjective health assessments and active labor market participation of older men: evidence from a semiparametric binary choice model with nonadditive correlated individual-specific effects. *Review of Economics and Statistics*, 93(3):764–774, 2011.

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

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

[97] Whitney K Newey. Nonparametric instrumental variables estimation. *The American Economic Review*, 103(3):550–556, 2013.

[98] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, pages 1565–1578, 2003.

[99] Whitney K Newey, James L Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, pages 565–603, 1999.

[100] Serena Ng and Joris Pinkse. Nonparametric-two-step estimation of unknown regression functions when the regressors and the regression error are not independent. 1995.

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

[102] Frank Partnoy. Rethinking regulation of credit rating agencies: An institutional investor perspective. *Council of Institutional Investors, April*, pages 09–014, 2009.

[103] George E Pinches and Kent A Mingo. A multivariate analysis of industrial bond ratings. *The journal of Finance*, 28(1):1–18, 1973.

[104] Joris Pinkse. Nonparametric two-step regression estimation when regressors and error are dependent. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 289–300, 2000.

[105] James L Powell. Estimation of semiparametric models. *Handbook of econometrics*, 4: 2443–2521, 1994.

[106] 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.

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

[108] 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.

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

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

[111] Michael Smith and Robert Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

[112] Robert M Solow. Technical change and the aggregate production function. *The review of Economics and Statistics*, pages 312–320, 1957.

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

[114] Liangjun Su, Stefan Hoderlein, and Halbert White. Testing monotonicity in unobservables with panel data. In *V Cowles Conference*, 2010.

[115] 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.

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

[117] Hal R Varian. The nonparametric approach to production analysis. *Econometrica: Journal of the Econometric Society*, pages 579–597, 1984.

[118] HD Vinod and Aman Ullah. Flexible production function estimation by nonparametric kernel estimators. 1987.

[119] Lawrence J White. The credit rating industry: An industrial organization analysis. In *Ratings, rating agencies and the global financial system*, pages 41–63. Springer, 2002.

[120] Jeffrey M Wooldridge. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, 104(3):112–114, 2009.