Informational Content of Factor Structures in Simultaneous Binary Response Models*

Shakeeb Khan
Boston College

Arnaud Maurel
Duke University, NBER and IZA

Yichong Zhang
Singapore Management University

September 2019

Abstract

We study the informational content of factor structures in discrete triangular systems. Factor structures have been employed in a variety of settings in cross sectional and panel data models, and in this paper we formally quantify their identifying power in a bivariate system often employed in the treatment effects literature. Our main findings are that imposing a factor structure yields point identification of parameters of interest, such as the coefficient associated with the endogenous regressor in the outcome equation, under weaker assumptions than usually required in these systems. In particular, we show that an exclusion restriction, requiring an explanatory variable in the outcome equation to be excluded from the treatment equation, is no longer necessary for identification. Under such settings, we propose a rank estimator for both the factor loading and the causal effect parameter that are root-n consistent and asymptotically normal. The estimator's finite sample properties are evaluated through a simulation study. We also establish identification results in models with more general factor structures, that are characterized by nonparametric functional forms and multiple idiosyncratic shocks.

Keywords: Factor Structures, Discrete Choice, Causal Effects.

1 Introduction

Factor models (or structures) see widespread and increasing use in various areas of econometrics. This type of structure has been employed in a variety of settings in cross sectional, panel and time series models, and have proven to be a flexible way to model the behavior of and relationship between unobserved components of econometric models. The baseline idea behind factor models is to assume that the dependence across the unobservables is generated by a low-dimensional set of mutually independent random variables (or factors). The applied and theoretical research in econometrics employing factor structures is extensive. In particular, these models are often used in the treatment

*We are thankful to seminar participants at Arizona State University, Emory, Michigan State, Shanghai University of Finance and Economics, University of Arizona, and conference participants at the 2015 SEA meetings for helpful comments. Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant no. MOE2018-T2-2-169 and the Lee Kong Chian fellowship.

effect literature as a way to identify the joint distribution of potential outcomes from the marginals, and then recover the distribution of treatment effects from this joint distribution.[1] Factor models have been used in a number of different contexts in applied microeconomics. Notably, factor models have been used in the context of earnings dynamics (Abowd and Card 1989, Bonhomme and Robin 2010), estimation of returns to schooling and work experiences (Ashworth, Hotz, Maurel, and Ransom 2017), as well as cognitive and non-cognitive skill production technology (Cunha, Heckman, and Schennach 2010), among others. All of these papers, with the notable exception of Cunha, Heckman, and Schennach (2010), rely on linear factor models where the unobservables are assumed to be given by the sum of a linear combination of mutually independent factors and an idiosyncratic shock.

In this paper we bring together the literature on factor models with the literature on the identification and estimation of triangular binary choice models (Chesher 2005, Vytlacil and Yildiz 2007, Shaikh and Vytlacil 2011, Han and Vytlacil 2017) by exploring the *informational content* of factor structures in this class of models. Focusing on this class can be well motivated from both an empirical and theoretical perspective. From the former, many treatment effect models fit into this framework as treatment is typically a binary and endogenous variable in the system, whose effect on outcomes is often a parameter the econometrician wishes to conduct inference on. From a theoretical perspective, inference on this type of system can be complicated, if not impossible without strong parametric assumptions, which may not be reflected in the observed data. A semiparametric approach to these models, while desirable from a theoretical point of view because of its generality, often fails to achieve identification of parameter, or at best only do so in sparse regions of the data, thus making inference impractical in practice. In this context, imposing a factor structure may be a useful "in-between" setting, which, at the very least, can be used to gauge the sensitivity of the parametric approach to their stringent assumptions.

We impose a particular factor structure to the two unobservables in this system and explore the informational content of this assumption. Specifically, we assume that the unobservables from the treatment equation (V) and the outcome equation (U) are related through the following factor model:

$$U = \gamma_0 V + \Pi$$

where Π is an unobserved random variable assumed to be distributed independently of V.[2] Our main finding in this case is that there is indeed informational content of factor structures in the sense that, in contrast to prior literature - notably Vytlacil and Yildiz (2007) - one no longer requires an additional exclusion restriction nor the strong support conditions that are needed for

[1] See Abbring and Heckman (2007) for a extensive discussion of factor structures and prior studies using these models in the context of treatment effect estimation.

[2] While this is our baseline specification, we also examine the informational content of more general factor structures involving nonparametric relationships between unobservables or multiple idiosyncratic errors.

identification in these models without the factor structure. Importantly, our identification results are constructive and translate directly into a rank based estimator of the coefficient associated with the binary endogenous variable.

The rest of the paper is organized as follows. In the next section we formally describe the triangular system with our factor structure, and discuss our main identification results for the parameters of interest in this model. Section 3 then proposes the estimation procedure and establishes its asymptotic properties. Section 4 explores identification in more complicated factor structure models which involve nonparametric relationships between unobservables or multiple idiosyncratic errors. Section 5 demonstrates the finite sample properties of the estimator proposed in this paper through Monte Carlo simulations. Section 6 concludes. The Supplementary Material collects the proofs of our results.

2 Triangular Binary Model with Factor Structure

In this section we consider the identification of the following factor structure model:

$$Y_1 = \mathbf{1}\{Z_1'\lambda_0 + Z_3'\beta_0 + \alpha_0 Y_2 - U > 0\}. \tag{2.1}$$

Turning to the model for the endogenous regressor, the binary endogenous variable Y_2 is assumed to be determined by the following reduced-form model:

$$Y_2 = \mathbf{1}\{Z'\delta_0 - V > 0\}, \tag{2.2}$$

where $Z \equiv (Z_1, Z_2)$ is the vector of "instruments" and (U, V) is a pair of random shocks. The subcomponent Z_2, Z_3 provides the exclusion restrictions in the model and is required to be non-degenerate conditional on $Z_1'\lambda_0 + Z_3'\beta_0$. We assume that the error terms U and V are jointly independent of (Z_1, Z_2, Z_3). The endogeneity of Y_2 in (2.1) arises when U and V are not independent.

The above model, or minor variations of it, have often been considered in the recent literature. See for example, Vytlacil and Yildiz (2007), Abrevaya, Hausman, and Khan (2010), Klein, Shan, and Vella (2015), Vuong and Xu (2017), Khan and Nekipelov (2018) and references therein. A key parameter of interest in our paper and in the rest of the literature is α_0. From the outset it is important to note that we provide conditions under which the parameters of interest are point-identified. As such, our analysis complements alternative partial-identification approaches that have been proposed for this type of triangular binary model. See, in particular, Chiburis (2010), Shaikh and Vytlacil (2011), and Mourifié (2015). As discussed in the aforementioned papers, this parameter is difficult, if not impossible to identify and estimate without imposing parametric

restrictions on the unobserved variables in the model, (U, V). Such parametric restrictions, such as the often assumed bivariate normality assumption, are not robust to misspecification in the sense that any estimator of α_0 based on these conditions will be inconsistent if (U, V) have a different bivariate distribution.

The established difficulty of identifying α_0 in semi parametric, i.e., "distribution free" models, and the sensitivity of its identification to misspecification in parametric models is what motivates the factor structure we add to the above model in this paper. Specifically, to allow for endogeneity in the form of possible correlation between U, V, we augment the model and add the following equation:

$$U = \gamma_0 V + \Pi \tag{2.3}$$

where Π is an unobserved random variable, assumed to be distributed independently of (V, Z_1, Z_2, Z_3), and γ_0 is an additional unknown scalar parameter. Importantly, this type of factor structure always holds when the residuals of both equations are jointly normally distributed. Furthermore, this specification corresponds to the type of structure used in Independent Component Analysis (ICA), where V and Π are two mutually independent factors. This method has found many applications in various fields, including signal processing and image extraction; applications in economics include e.g., Hyvärinen and Oja (2000), Moneta, Hoyer, and Coad (2013) and Gourieroux, Monfort, and Renne (2017).

Our aim will be to first explore identification for the parameters $(\alpha_0, \delta_0, \gamma_0, \beta_0, \lambda_0)$ under standard nonparametric regularity conditions on (V, Π). It is interesting to note that our approach, which consists in adding more structure to the fully semiparametric triangular binary system and quantify the identifying power of the added structure is, in one sense, the reverse approach of generalizing the fully parametric model. Such an approach has been taken recently in Han and Vytlacil (2017), who begin with a bivariate Probit model, and generalize it with the introduction of a class of one parameter copulas, providing conditions such that identification can still be obtained.[3]

Our linear factor structure and the one-parameter copula model considered in Han and Vytlacil (2017) are not nested by each other. Based on the factor structure, we can recover F_Π, the distribution of Π, as a function of (F_U, F_V, γ_0) by deconvolution. Then we can write the copula of (U, V) as

$$F_{U,V}(F_U^{-1}(u), F_V^{-1}(v)) = \int_{-\infty}^{F_V^{-1}(v)} F_\Pi(F_U^{-1}(u) - \gamma_0 w; F_U, F_V, \gamma_0) f_V(w) dw = C(u, v; F_U, F_V, \gamma_0).$$

The copula depends not only on γ_0 but also on two infinite dimensional parameter (F_U, F_V). Thus, unlike Han and Vytlacil (2017), the factor structure cannot be characterized by a one-parameter copula. In addition, in order to achieve identification, Han and Vytlacil (2017) first

[3]See also recent work by Han and Lee (2018) who study semiparametric estimation and inference in the framework considered by Han and Vytlacil (2017).

nonparametrically identify the two marginals by assuming the existence of a full support regressor that is common to both equations. In contrast, our approach does not rely on the existence of a full support common regressor. Under the factor structure assumed in this paper, we bypass the nonparametric identification of the marginals as a whole and directly consider the identification of the structural parameters. It follows that our model cannot be nested by the one-parameter copula model considered by Han and Vytlacil (2017). On the other hand, there exists one-parameter copula models that cannot be decomposed into linear factor structures.[4] This implies that our model does not nest Han and Vytlacil (2017) either.

To simplify the exposition of our strategy, in this and the following sections we will focus exclusively on the parameters α_0, γ_0 and denote the linear indices by $X_1 \equiv Z_1'\lambda_0 + Z_3'\beta_0$ and $X \equiv Z'\delta_0$, where $Z = (Z_1, Z_2)$. In particular, we treat δ_0 as known. In practice, δ_0 can be identified and consistently estimated in a first step using a semi-parametric single index estimator such as the one proposed by Klein and Spady (1993). In addition, at the end of this section, we note that we can identify λ_0 and β_0 simultaneously with α_0. Then (2.1) and (2.2) are simplified to

$$Y_1 = \mathbf{1}\{X_1 + \alpha_0 Y_2 - U > 0\} \tag{2.4}$$

and

$$Y_2 = \mathbf{1}\{X - V > 0\}. \tag{2.5}$$

Our proof will be based on the Assumptions **A1**-**A5** we state here:

A1 The parameter $\theta_0 \equiv (\alpha_0, \gamma_0)$ is an element of a compact subset of $\Re^2$.

A2 The vector of unobserved variables, (U, V, Π) is continuously distributed with support on R^3 and independently distributed of the vector (Z_1, Z_2, Z_3). Furthermore, we assume the unobserved random variables Π, V are distributed independently of each other.

A3 X is continuously distributed with absolute continuous density w.r.t. Lebesgue measure. The density is bounded and bounded away from zero on any compact subset of its support.

A4 For any constant c, $P(X - \tilde{X} = c | X_1 + \alpha_0 - \gamma_0 X = \tilde{X}_1 - \gamma_0 \tilde{X}) < 1$, where $(\tilde{X}, \tilde{X}_1)$ are an independent copy of (X, X_1).

A5 $\text{Supp}(X_1 + \alpha_0 - \gamma_0 X) \cap \text{Supp}(X_1 - \gamma_0 X) \neq \emptyset$.

Before turning to our main identification result, a couple of remarks are in order.

[4]For instance, suppose that (U, V) has a Gaussian copula with correlation ρ, and that the marginal distributions of U and V are uniform $[0, 1]$. It then follows that, denoting by $\Phi(.)$ the standard normal cdf., $\left(\Phi^{-1}(U), \Phi^{-1}(V)\right)$ is bivariate normal with correlation ρ, which in turn yields the following non-linear relationship between U and V: $U = \Phi\left(\rho\Phi^{-1}(V) + W\right)$, where W is normally distributed and independent from V.

Remark 2.1. *Assumption A2 is standard in this literature in both the unobservables U, V as well as the independence between Π and V. References for the former (instruments independent of unobservables), can be found in Abrevaya, Hausman, and Khan (2010), Vytlacil and Yildiz (2007), Klein, Shan, and Vella (2015), Khan and Nekipelov (2018). For the latter (Π independent of V), see, e.g. Bai and Ng (2002) and Carneiro, Hansen, and Heckman (2003).*

Remark 2.2. *Assumption A3 requires the instrumental variable Z_2 in the selection equation to be continuously distributed, which is often required in models with discrete outcomes.*

Remark 2.3. *Assumption A4 requires some variation of X_1. In particular, X_1 cannot be a constant but is allowed to be discrete.*

Remark 2.4. *A sufficient condition for Assumption A5 is that the length of support of $X_1 - \gamma_0 X$ is less than or equal to α_0, which is a form of parameter space constraint. It is analogous to that imposed in Vytlacil and Yildiz (2007), but distinct in important ways. Specifically, the length of the support of the instrument Z_2, and thus X, now helps in the identification of α_0. This is natural, as a purpose of the instrument Z_2 should benefit in the identification of the parameters of the outcome equation similar to standard IV approaches for the linear model. Another aspect of Assumption A5 is it imposes no constraints directly on the variables in the outcome equation. Specifically, β_0 can be zero and X can be discrete, yet we still can attain identification. Therefore, the factor structure replaces the need for a continuous exclusion variable in the outcome equation, distinguishing our point identification result from those in Vytlacil and Yildiz (2007) and Vuong and Xu (2017).*[5]

We now turn to Theorem 2.1 below, which concludes that under our stated conditions and our factor structure we can attain point identification.

Theorem 2.1. *Under Assumptions A1-A5, θ_0 is point identified.*

Let $P^{ij}(x_1, x) = Prob(Y_1 = i, Y_2 = j | X_1 = x_1, X = x)$ and $\partial_2 P^{ij}(x_1, x)$ denote the choice probability and the the derivative of the ij-choice probability with respect to the second argument, respectively. Then, both are identified from data. The proof of Theorem 2.1, which is reported in Section A in the Supplementary Appendix, relies on the fact that, for two pairs of observations (X_1, X) and $(\tilde{X}_1, \tilde{X})$,

$$\partial_2 P^{11}(X_1, X)/f_V(X) + \partial_2 P^{10}(\tilde{X}_1, \tilde{X})/f_V(\tilde{X}) = 0 \iff X_1 + \alpha_0 - \gamma_0 X - (\tilde{X}_1 - \gamma_0 \tilde{X}) = 0, \quad (2.6)$$

where $f_V(\cdot)$ is the PDF of V, which can be identified over the support of X. (2.6) shows the variation in X can be used to identify α_0, even when X_1 is discrete.

In addition, recall that $X_1 = Z_1'\lambda_0 + Z_3'\beta_0$. Suppose X_1, and thus, all the elements of Z_1 and Z_3 are discrete, then it is impossible to point identify λ_0 and β_0 by simply using the outcome equation.

[5] In Section F of the Supplementary Appendix we extend these arguments to the case where we do not attain point identification. Specifically, we show that the factor structure enables sharper bounds for α_0 than bivariate models without factor structures.

Further assume we can identify δ_0, and thus, can treat X as an observable. Then, we can identify the choice probability

$$P^{ij}(z_1, z_3, x) = Prob(Y_1 = i, Y_2 = j | Z_1 = z_1, Z_3 = z_3, X = x)$$

and its derivative w.r.t. x, i.e., $\partial_2 P^{ij}(z_1, z_3, x)$. Then, by the same argument, we can show that

$$\begin{aligned} &\partial_2 P^{11}(Z_1, Z_3, X)/f_V(X) + \partial_2 P^{10}\tilde{(}Z_1, \tilde{Z}_3, \tilde{X})/f_V(\tilde{X}) = 0 \\ \iff &Z_1'\lambda_0 + Z_3'\beta_0 + \alpha_0 - \gamma_0 X - (\tilde{Z}_1'\lambda_0 + \tilde{Z}_3'\beta_0 - \gamma_0\tilde{X}) = 0. \end{aligned}$$

Then, given sufficient variation in X, we can identify (λ_0, β_0) along with (α_0, γ_0) even when all elements of Z_1 and Z_3 are discrete.[6]

An important takeaway from this result is that imposing our factor structure yields point-identification under weaker support condition when compared to the existing literature, and does not require the second exclusion restriction either. In particular, our results yield point-identification of the parameters of interest even in situations where all of the regressors from the outcome equation are discrete. Interestingly, this indicates that, from the selection equation combined with the factor structure that we impose here, we can overturn the non-identification result of Bierens and Hartog (1988) which would apply to the outcome equation alone.

3 Estimation and Asymptotic Properties

The previous section established a point identification result. The identification result is constructive in the sense that it motivates an estimator for for the parameters of interest which we describe in detail here.

As we did in Section 2, to simplify exposition, in the following we focus exclusively on the parameters α_0, γ_0. Recall the choice probabilities $P^{ij}(x_1, x) = Prob(Y_1 = i, Y_2 = j | X_1 = x_1, X = x)$ and its second derivative $\partial_2 P^{ij}(x_1, x)$, which can be estimated as we describe below. Another function needed for our identification result is the density function of the unobserved term V, denoted by $f_V(\cdot)$. This is also unknown, but from the structure of our model can be recovered from the derivative with respect to the instrument Z of $E[Y_2|Z]$, and hence is estimable from the data. Note that the proof of Theorem 2.1 shows that the sign of the index evaluated at two *different* regressor values, which we denote here by X and $\tilde{X}$ is determined by the choice probabilities via

$$\partial_2 P^{11}(X_1, X)/f_V(X) + \partial_2 P^{10}(\tilde{X}_1, \tilde{X})/f_V(\tilde{X}) \geq 0 \quad \iff X_1 + \alpha - \gamma X - (\tilde{X}_1 - \gamma\tilde{X}) \geq 0.$$

[6] An alternative approach to identifying this parameter can be found in Lewbel (2000). In his approach a second equation to model the endogenous variable is not needed, nor is the factor structure we impose. However, he imposes a strong support condition on a variable like Z_3 requiring that it exceeds the length of the unobservable U. As explained in Khan and Tamer (2010), such an approach precludes even bounding α_0 if the support condition on Z_3 is not satisfied.

This motivates us to use maximum rank correlation estimator proposed by Han (1987).

Implementation requires further details to pay attention to. The unknown choice probabilities, their derivatives, and the density of V will be estimated using nonparametric methods, and for this we adopt locally linear methods as they are particularly well suited for estimating derivatives of functions.

With functions and their derivatives estimated in the first stage of our procedure, the second stage plugs in these estimated values into an objective function to be optimized. Specifically, letting $\hat{\theta}$ denote $(\hat{\alpha}, \hat{\gamma})$, our estimator is of the form:

$$\hat{\theta} = \arg\max_{\theta} Q_{n,2}(\theta) \equiv \sum_{i \neq j} \hat{g}_{i,j}(\theta) \tag{3.1}$$

in which

$$\begin{aligned} \hat{g}_{i,j}(\theta) &= [\mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \\ &+ \mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\}], \end{aligned}$$

with

$$\Phi(x_1, x, \tilde{x}_1, \tilde{x}; \theta) = x_1 + \alpha - \gamma x - (\tilde{x}_1 - \gamma \tilde{x}).$$

We note that this estimator falls into the class of those which optimize a nonsmooth U-process involving components estimated nonparametrically in a preliminary stage.[7] Examples of other estimators in this class can be found in Khan (2001), Abrevaya, Hausman, and Khan (2010), Jochmans (2013), Chen, Khan, and Tang (2016), and our approach to deriving the limiting distribution theory of our estimator will follow along the steps used in those papers. Our limiting distribution theory for this estimator is based on the following regularity conditions:

RK1 θ_0 lies in the interior of Θ, a compact subset of R^2.

RK2 The index X is continuously distributed with support on the real line, and has a density function which is twice continuously differentiable.

RK3 (Order of smoothness of probability functions and regressor density functions) The functions $P^{k,l,r}(\cdot)$ and $f_{X_1,X}(\cdot\cdot)$ (the density function of the random vector (X_1, X)) are continuously differentiable of order p_2, where $p_2 > 5$.

[7] An alternative estimation procedure could be based on the exact relationship in (2.6). Note the equality on the left-hand side of (2.6) is a function of the data alone and not the unknown parameters. The right-hand side equality can then be regarded as a moment condition to estimate the unknown parameters. We describe this estimator and derive its asymptotic properties in the Online Supplement to the paper. While the two estimation approaches will have similar asymptotic properties (root-n consistent, asymptotically normal), we prefer the rank estimator in (3.1) which involves fewer tuning parameters. Furthermore rank type estimators in general are more robust to certain types of misspecification, as pointed out in Khan and Tamer (2018).

RK4 (First stage kernel function conditions) $K(\cdot)$, used to estimate the choice probabilities and their derivatives is an even function, integrating to 1 and is of order p_2 satisfying $p_2 > 5$.

RK5 (Rate condition on first stage bandwidth sequence) The first stage bandwidth sequence H_n used in the nonparametric estimator of the choice probability functions and their derivatives satisfies $\sqrt{n}H_n^{p_2-1} \to 0$ and $n^{-1/4}H_n^{-1} \to 0$.

Based on these conditions, we have the following theorem, whose proof is in Section B of the Supplementary Appendix which characterizes the rate of convergence and asymptotic distribution of the proposed estimator:

Theorem 3.1. *Under Assumptions* **RK1-RK5**,

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, V^{-1}\Delta V^{-1}) \tag{3.2}$$

where the forms of the Hessian term V and outer score term Δ are described in detail in Section B of the Supplementary Appendix.

4 More General Factor Structures

Up until now we have proposed identification and estimation results for a triangular system with a particular factor structure. A disadvantage of this structure is that it is restrictive in two ways. One is that it is a model with only one idiosyncratic shock (Π). The other is the linear in parameter relationship between the two unobserved components, which leaves open the possibility of misspecification.

4.1 Nonparametric Factor Model

Consider the following relationship between unobserved components:

$$U = g_0(V) + \tilde{\Pi} \tag{4.1}$$

where $\tilde{\Pi}$ is an unobserved random variable assumed to be distributed independently of V and all instruments. $g_0(\cdot)$ is an unknown function assumed to satisfy standard smoothness conditions. Again, the parameter of interest is α_0, but now the unknown nuisance parameter in the factor equation is infinite dimensional. Now our approach is to replace the vector X with a series of basis functions of X, such as, for example orthonormal polynomials, in X. Those basis functions are meant to serve as an approximation of $g_0(\cdot)$. With that replacement, we carry exactly as before, except now instead of estimating a kernel weighted linear regression model it will be a kernel

weighted semi linear regression model as in, e.g., Robinson (1988). Section C in the Supplementary Appendix provides details of how to construct such an estimator and outlines its large sample properties.

4.2 Model with Two Idiosyncratic Shocks and a Bounded Common Factor

We express this model as:

$$\begin{aligned} Y_1 &= \mathbf{1}\{X_1 + \alpha_0 Y_2 - U \geq 0\} \\ Y_2 &= \mathbf{1}\{X - V \geq 0\}, \end{aligned} \tag{4.2}$$

where $U = \gamma_0 W + \eta_1$, $V = W + \eta_2$, and (W, η_1, η_2) are mutually independent. First we consider the case $\gamma_0 = 1$ and X_1 is binary, because even in this context, for the baseline case with one idiosyncratic shock, we can identify α_0. But identification of α_0 becomes more difficult in this model, as established in the following theorem

Theorem 4.1. *Suppose* (4.2) *holds,* γ_0 *is known to be one,* X_1 *is binary, and* W *has a bounded support* $[-b, -a]$ *such that* $0.5 > b - a$ *and* $1 - (b - a) > \alpha_0 > b - a$*, then* α_0 *is* **not** *point identified.*

This nonidentification result motivates imposing additional structure on W, and we consider the following model

B1 $U = \gamma_0 W + \eta_1 \quad$ and $\quad V = \sigma_0 W + \eta_2$.

B2 W is standard normally distributed.

B3 W, η_1 and η_2 are mutually independent.

B4 X has full support.

B5 Denote the density of V as f_V, then f_V does not have a Gaussian component in the sense that

$$f_V \in \mathcal{G} = \{g \text{ is a density on } \Re \text{ s.t. : } g = g' * \phi_\sigma \text{ for some density } g' \text{ implies that } \sigma = 0\},$$

where ϕ_σ is the density for a normal distribution with zero mean and σ^2 variance.

Assumption **B5** effectively assumes that the distribution of η_2 has tail properties different from those of a normal distribution. This type of assumption is made in the deconvolution literature as it is necessary for identification of the target density when the error distribution is not completely known- see, e.g., Butucea and Matias (2005).[8] The importance of non-normality in factor models goes back to Geary (1942) and Reiersol (1950), who have shown that factor loadings are identified in a linear measurement error model if the factor is not Gaussian.

[8] In fact, based on the results in Butucea and Matias (2005), W can belong to a more general class of known distributions. Furthermore, we note that if σ_0 is known, then Assumption **B5** is not necessary.

Theorem 4.2. *If Assumptions **B1**–**B5** hold, then σ_0, γ_0 and α_0 are identified.*

Note that this identification result does not require any variation from X_1, which is in spirit close to the one-factor model in our paper and is different from the identification result in Vytlacil and Yildiz (2007). We also note that this result does not contradict the counterexample in the paper. In the counterexample, we only assume that we know the support of W is bounded. Here we assume that the full density of W, and thus, the support of W is known.

5 Finite Sample Properties

In this section we explore the finite sample properties of the proposed estimation procedure via a simulation study. We will also see how sensitive the performance of the proposed estimator is to the factor structure assumption. As a base comparison, we also report results for the estimator proposed in Vytlacil and Yildiz (2007) to see how sensitive it is to their second instrument restriction.

Our data are simulated from base models of the form

$$Y_1 = \mathbf{1}\{X_1 + \alpha_0 Y_2 - U \geq 0\} \tag{5.1}$$

$$Y_2 = \mathbf{1}\{X - V > 0\} \tag{5.2}$$

where X_1 is binary with success probability 0.6, X has marginal distribution $\mathcal{N}(0,1)$, X_1 and X are mutually independent, $(X_1, X) \perp (V, \Pi)$, $U = \gamma_0 V + \Pi$. (V, Π) are distributed independently of each other, where V is distributed following a standard normal distribution, and Π is distributed either standard normal, Laplace, or $T(3)$. The parameters $(\alpha_0, \gamma_0) = (-0.25, 1.2)$ or $(0.5, 1.2)$.

Since X_1 is discrete, Vytlacil and Yildiz (2007)'s identification condition does not hold. However, the identification condition in this paper becomes

$$|\alpha| \leq \text{length of the support of } X,$$

which holds.

For each choice of sample size $n = 100, 200, 400, 800, 1,600$, we simulate 280 samples and report the bias, standard deviation (std), root mean squared error (RMSE), and median absolute deviation (MAD) for both Vytlacil and Yildiz (2007)'s estimator (VY) and ours (KMZ). For implementation, we use the second order local polynomial along with Gaussian kernels to nonparametrically estimate the $\partial_2 P^{11}(x_1, x)$ and $\partial_2 P^{10}(x_1, x)$. The bandwidth we use is $h_1 = \sigma_x N^{-1/7}$ where σ_x is the standard deviation of X. $f_V(x)$ is nonparametrically estimated using local linear estimator with the tuning parameter $h_2 = \sigma_x N^{-1/6}$.

As results from the table indicate, the finite sample performance of our estimator generally agrees with the asymptotic theory. The RMSE for the estimator proposed here is decreasing as the sample size increases, as one could expect given the consistency property of our estimator. Besides, the decay rate of the RMSE and MAD is about $\sqrt{2}$ when $n \geq 400$ as sample sizes doubles, in line with the parametric rate of convergence of our estimator.

The Vytlacil and Yildiz (2007) estimator, which does not exploit the factor structure demonstrates inconsistency for certain parameter values, as indicated by the bias and median bias not shrinking with the sample size. In addition, the RMSE and MAD do not appear to decline at all, which also suggests that Vytlacil and Yildiz (2007)'s estimator is inconsistent in these designs.[9]

Table 1: Normal V, $\alpha = 0.5$

Π	Normal						Laplace						T(3)					
	kmz			vy			kmz			vy			kmz			vy		
N	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD
100	-0.026	0.665	0.660	-0.246	0.658	0.500	0.032	0.634	0.560	-0.293	0.658	0.500	0.010	0.676	0.665	-0.225	0.662	0.500
200	0.004	0.591	0.475	-0.329	0.633	0.500	-0.015	0.568	0.400	-0.336	0.612	0.500	-0.003	0.616	0.495	-0.279	0.629	0.500
400	0.005	0.483	0.365	-0.341	0.573	0.500	0.030	0.459	0.310	-0.323	0.559	0.500	0.018	0.542	0.405	-0.314	0.589	0.500
800	0.065	0.456	0.300	-0.348	0.544	0.500	0.096	0.391	0.250	-0.357	0.511	0.500	0.046	0.462	0.295	-0.346	0.552	0.500
1,600	0.040	0.321	0.195	-0.413	0.503	0.500	0.017	0.294	0.190	-0.450	0.506	0.500	0.034	0.371	0.240	-0.368	0.506	0.500

Table 2: Normal V, $\alpha = -0.25$

Π	Normal						Laplace						T(3)					
	kmz			vy			kmz			vy			kmz			vy		
N	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD
100	-0.088	0.650	0.555	-0.466	0.710	0.750	0.092	0.614	0.530	-0.358	0.650	0.750	0.004	0.619	0.505	-0.430	0.681	0.750
200	-0.035	0.599	0.420	-0.446	0.681	0.750	0.012	0.552	0.385	-0.485	0.689	0.750	-0.008	0.583	0.425	-0.463	0.687	0.750
400	-0.016	0.467	0.325	-0.487	0.668	0.750	-0.010	0.388	0.200	-0.552	0.686	0.750	-0.003	0.496	0.340	-0.489	0.675	0.750
800	-0.028	0.324	0.165	-0.591	0.697	0.750	0.006	0.279	0.180	-0.599	0.701	0.750	0.032	0.399	0.230	-0.533	0.682	0.750
1,600	-0.006	0.244	0.150	-0.654	0.718	0.750	-0.028	0.204	0.130	-0.714	0.738	0.750	-0.021	0.279	0.190	-0.629	0.710	0.750

In the following, we also consider DGPs such that the one-factor model does not hold but the identification assumption in Vytlacil and Yildiz (2007) does. In this case, our simulation results show that while, as expected, the estimator VY is still valid, our estimator still performs reasonably well. Interestingly, this offers suggestive evidence that our estimator is robust to some degree of misspecification. As such, these results complement previous work highlighting the robustness of rank type estimators to misspecification (Khan and Tamer 2018).

The outcome and selection equations are the same as (2.1) and (2.2), respectively. Then,

DGP 1 : (X_1, X) is jointly standard normally distributed. Let (e_1, e_2) jointly Laplace distributed

[9]Because X_1 is binary, the Vytlacil and Yildiz (2007) estimator can only take 3 possible values: 0, -1 or 1. In particular, when $\alpha = 0.5$, in most of the replications, the estimator takes values 0 or 1. When $\alpha = -0.25$, in most of the replications, the estimator takes value -1. In both of these cases, the MAD remains constant over the different sample sizes.

with mean zero and variance-covariance matrix $\Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$, e_3 and e_4 are uniformly distributed on $(0,1)$, independent of each other, and independent of (e_1, e_2), $V = e_1 + e_3 - 0.5$, $U = e_2 + e_4 - 0.5$, and $\alpha = -0.25$.

DGP 2 : (X_1, X) are the same as above, $V = e_1 + e_2 - 0.5$, and $V = e_1 + e_3 - 0.5$, where e_1 is standard normally distributed, (e_2, e_3) are uniformly distributed on $(0,1)$, (e_1, e_2, e_3) are mutually independent, and $\alpha = -0.25$.

DGP 3 : (X_1, X) are the same as above, $V = \frac{\exp(e_1+e_2-0.5)-1}{4}$, $U = \frac{\exp(e_1+e_3-0.5)-1}{4}$, (e_1, e_2, e_3) are defined as above, and $\alpha = -0.5$.

DGP 4 : (X_1, X) are the same as above, V is Laplace distributed with mean zero and standard derivation 0.5, $U = V + V' - 0.5$, where V' is uniform distributed on $(0,1)$ and is independent of V, and $\alpha = -0.25$.

For DGPs 1, 2, and 4, when computing $\partial_2 P^{11}(x_1, x)$ and $\partial_2 P^{10}(x_1, x)$, we use bandwidths $h_1 = \sigma_{x1} N^{-1/7}$ and $h = \sigma_x N^{-1/7}$ for variables X_1 and X, respectively, where σ_{x1} and σ_x are the standard errors of X_1 and X, respectively. To estimate the density $f_V(x)$, we use bandwidth $h_2 = \sigma_x N^{-1/6}$. For DGP 3, we use $h_1 = h_2 = h = \sigma_{x1} N^{-1/5}$. In all simulations, we use 280 replications.

Table 3: Alternative DGPs

	DGP 1						DGP 2					
	kmz			vy			kmz			vy		
N	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD	Bias	RMSE	MAD
100	-0.065	0.678	0.600	-0.055	0.666	0.535	-0.058	0.621	0.505	-0.05	0.621	0.470
200	-0.118	0.543	0.370	-0.080	0.497	0.320	-0.122	0.523	0.350	-0.097	0.495	0.350
400	-0.117	0.413	0.280	-0.071	0.378	0.245	-0.062	0.335	0.215	-0.033	0.316	0.220
800	-0.102	0.287	0.170	-0.062	0.243	0.160	-0.031	0.242	0.150	-0.008	0.215	0.150
1,600	-0.071	0.193	0.140	-0.035	0.155	0.100	-0.038	0.167	0.100	-0.031	0.158	0.100
	DGP 3						DGP 4					
100	-0.012	0.583	0.480	-0.015	0.565	0.430	-0.057	0.401	0.240	-0.066	0.422	0.240
200	-0.061	0.425	0.275	-0.068	0.399	0.270	-0.041	0.282	0.180	-0.049	0.263	0.145
400	-0.041	0.259	0.170	-0.042	0.237	0.155	-0.062	0.184	0.135	-0.047	0.186	0.120
800	-0.061	0.219	0.140	-0.047	0.182	0.120	-0.029	0.119	0.080	-0.034	0.115	0.070
1,600	-0.038	0.130	0.080	-0.035	0.119	0.080	-0.024	0.090	0.060	-0.022	0.086	0.070

In the first three DGPs, we see that VY's estimator has better performance in terms of both bias and MSE. On the other hand, although the models do not have a factor structure, our estimator still performs reasonably well. In the last DGP, both VY's and our estimator are consistent and exhibit similar finite sample performance.

6 Conclusions

In this paper we explored the identifying power of factor structures in discrete simultaneous systems. We found that for a binary-binary system the factor structure we considered did indeed add informational content. Specifically, it enabled the relaxation of both the exclusion and support conditions typically employed in the identification of these models. As we then demonstrated factor structures then enabled the regular identification of parameters of interest, and we proposed a new rank based estimation procedure that converged at a parametric rate with a limiting normal distribution. Finite sample properties of the estimator were demonstrated through simulation studies.

The work here opens areas for future research. The factor structure we assume could be imposed in more general models. For example, nontriangular discrete systems have shown to be an effective way to model entry games in the empirical industrial organization literature- see, for example Tamer (2003). However, as shown in Khan and Nekipelov (2018), identification of structural parameters in these models can be even more challenging than for the triangular model considered in this paper. It would be useful to determine if factor structures on the unobservables could alleviate this problem. We leave this open question to future work.

Supplement to "Informational Content of Factor Structures in Simultaneous Binary Response Models"

Abstract

This paper gathers the supplementary material to the original paper. Section A proves Theorem 2.1. Section B establishes the asymptotic distribution for the rank estimator. Section C describes an estimator for the case where we have a nonparametric factor structure. Sections D and E prove Theorems 4.1 and 4.2, respectively. Section F discuss the identification power of the factor structure. Section G contains the proof of the statement in Section F.

A Proof of Theorem 2.1

Proof: Note that

$$P^{11}(x_1, x) = \int_{-\infty}^{x} F_\Pi(x_1 + \alpha_0 - \gamma_0 v) f_V(v) dv$$
$$P^{10}(\tilde{x}_1, \tilde{x}) = \int_{\tilde{x}}^{+\infty} F_\Pi(\tilde{x}_1 - \gamma_0 v) f_V(v) dv.$$

Taking derivatives w.r.t. the second argument of the LHS function, we obtain

$$\partial_2 P^{11}(x_1, x)/f_V(x) = F_\Pi(x_1 + \alpha_0 - \gamma_0 x)$$
$$-\partial_2 P^{10}(\tilde{x}_1, \tilde{x})/f_V(\tilde{x}) = F_\Pi(\tilde{x}_1 - \gamma_0 \tilde{x}).$$

By Assumption A5, we know that there exists pairs $(x_1^{(1)}, x^{(1)})$ and $(\tilde{x}_1^{(1)}, \tilde{x}^{(1)})$ in $\text{Supp}(X_1, X)$ such that

$$x_1^{(1)} + \alpha_0 - \gamma_0 x^{(1)} = \tilde{x}_1^{(1)} - \gamma_0 \tilde{x}^{(1)}.$$

These pairs can be identified from data by the fact that

$$\partial_2 P^{11}(x_1^{(1)}, x^{(1)})/f_V(x^{(1)}) + \partial_2 P^{10}(\tilde{x}_1^{(1)}, \tilde{x}^{(1)})/f_V(\tilde{x}^{(1)}) = 0.$$

By Assumption A4, there exists at least another pair $(x_1^{(2)}, x^{(2)})$ and $(\tilde{x}_1^{(2)}, \tilde{x}^{(2)})$ in $\text{Supp}(X_1, X)$ such that

$$x_1^{(2)} + \alpha_0 - \gamma_0 x^{(2)} = \tilde{x}_1^{(2)} - \gamma_0 \tilde{x}^{(2)}, \text{ and } x^{(2)} - \tilde{x}^{(2)} \neq x^{(1)} - \tilde{x}^{(1)}.$$

So we have a two equation system

$$\alpha_0 - \gamma_0(x^{(1)} - \tilde{x}^{(1)}) = \tilde{x}_1^{(1)} - x_1^{(1)}$$
$$\alpha_0 - \gamma_0(x^{(2)} - \tilde{x}^{(2)}) = \tilde{x}_1^{(2)} - x_1^{(2)}.$$

Since $x^{(2)} - \tilde{x}^{(2)} \neq x^{(1)} - \tilde{x}^{(1)}$, the system of equations has a unique solution. This concludes the proof.

B Distribution Theory for the Rank Estimator

Recall we defined our two step rank estimator as follows: Letting $\hat{\theta}$ denote $(\hat{\alpha}, \hat{\gamma})$, our estimator is of the form:

$$\hat{\theta} = \arg\max_{\theta} \hat{Q}_{n,2}(\theta) \equiv \sum_{i \neq j} \hat{g}_{i,j}(\theta)$$

in which

$$\begin{aligned} \hat{g}_{i,j}(\theta) &= [\mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \\ &+ \mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\}], \end{aligned}$$

with

$$\Phi(x_1, x, \tilde{x}_1, \tilde{x}; \theta) = x_1 + \alpha - \gamma x - (\tilde{x}_1 - \gamma \tilde{x})$$

We note this estimator falls into the class of those which optimize a nonsmooth U-process.involving components estimated nonparametrically in a preliminary stage. Example of other estimators in this class can be found in Khan (2001), Abrevaya, Hausman, and Khan (2010), Jochmans (2013), Chen, Khan, and Tang (2016), and our approach to deriving the limiting distribution theory of our estimator will follow along the steps used in those papers. Our proof strategy will be based on deriving a quadratic approximation for the objective function $Q_{n,2}(\theta)$, in a way analogous to the method introduced in Sherman (1994b). Following Sherman (1994b), we will derive the asymptotic properties of $\hat{\theta}$ in three stages. We will first establish its consistency, then derive an intermediate rate (4^{th} root consistency), followed by establishing root-n consistency and asymptotic normality of the estimator. Our result are based on the following regularity conditions:

RK1 θ_0 lies in the interior of Θ, a compact subset of R^2.

RK2 The index X is continuously distributed with support on the real line, and has a density function which is twice continuously differentiable.

RK3 (Order of smoothness of probability functions and regressor density functions) The functions $P^{k,l,r}(\cdot)$ and $f_{X_1,X}(\cdot\cdot)$ (the density function of the random vector (X_1, X)) are continuously differentiable of order p_2, where $p_2 > 5$.

RK4 (First stage kernel function conditions) $K(\cdot)$, used to estimate the choice probabilities and their derivatives is an even function, integrating to 1 and is of order p_2 satisfying $p_2 > 5$.

RK5 (Rate condition on first stage bandwidth sequence) The first stage bandwidth sequence H_n used in the nonparametric estimator of the choice probability functions and their derivatives satisfies $\sqrt{n}H_n^{p_2-1} \to 0$ and $n^{-1/4}H_n^{-1} \to 0$.

We first show consistency of the rank estimator. To do so we first define the objective function $Q_{n,2}^{if}(\theta)$, defined as

$$Q_{n,2}^{if}(\theta) \equiv \sum_{i \neq j} g_{i,j}(\theta)$$

where

$$\begin{aligned} g_{i,j}(\theta) &= [\mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \\ &+ \mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\}], \end{aligned}$$

Since $g_{i,j}$ is bounded by 1 $\forall i, j$, and our random sampling assumption, we have for each θ,

$$Q_{n,2}^{if}(\theta) \xrightarrow{p} E[g_{i,j}(\theta)] \equiv \Gamma_0(\theta)$$

Furthermore, by Assumptions RK2, RK3 we can extend this result to converging uniformly over $\theta \in \Theta$ (see, e.g. Sherman (1994a), Sherman (1993).) $\Gamma_0(\theta)$ is continuous in θ by Assumptions RK2,RK3, and uniquely maximized at $\theta = \theta_0$ by our identification result in Theorem 2.1. Along with Assumption RK1, the infeasible estimator, defined as the maximizer of $Q_{n,2}^{if}(\theta)$ converges in probability to θ_0 by, for example Theorem 2.1 in Newey and McFadden (1994). To show consistency of the feasible estimator, where we first estimate the choice probability functions and their derivatives nonparametrically, we only now need to show the two objective functions converged to each other uniformly in $\theta \in \Theta$. Consistency of the first stage estimators follows from Assumptions **RK3-RK5**, see for example Henderson, Li, Parmeter, and Yao (2015). However, this does not immediately imply convergence of the difference in feasible and infeasible objective functions since the nonparametric estimators are inside indicator functions so the continuous mapping theorem does immediately not apply. Nonetheless the desired result can still be attained in one of two ways. One would be to replace indicator functions with smooth distribution functions in a fashion analogous to Horowitz (1992). This would have the disadvantage of introducing tuning parameters, but another approach would be to replace the indicator functions with their conditional expectations, and note that the conditional expectations are smooth functions using Assumption **RK2, RK3.**. To see why, let $\hat{m}(x_i)$ be a nonparametric estimator of a function $m(x_i)$, which is assumed to be smooth. We evaluate the plim of

$$I[\hat{m}(x_i) > 0] - I[m(x_i) > 0] = I[\hat{m}(x_i) > 0, m(x_i) < 0] - I[\hat{m}(x_i) < 0, m(x_i) > 0]$$

we show that the first term converges in probability to 0 as identical arguments can be used for the second term. Let $\varepsilon > 0$ be given; $P(I[\hat{m}(x_i) > 0, m(x_i) < 0] > \varepsilon) \leq E[I[\hat{m}(x_i) > 0, m(x_i) < 0]/\varepsilon$

by Markov's inequality. But the expectation in the numerator on the right hand side is

$$P(\hat{m}(x_i) > 0, m(x_i) < 0) = P(\hat{m}(x_i) > 0, m(x_i) \leq -\delta_n) + P(\hat{m}(x_i) > 0, m(x_i) \in (-\delta_n, 0))$$

where δ_n is a sequence of positive numbers converging to 0, at a slow rate, e.g.$(\log n^{-1})$. The first term on the right hand side is bounded above by

$$P(|\hat{m}(x_i) - m(x_i)| > \delta_n) \leq P(\|\hat{m}(\cdot) - m(\cdot)\| > \delta_n)$$

where the notation $\|\hat{m}(\cdot) - m(\cdot)\|$ above denotes the sup norm over x_i. The right hand side probability above will be sufficiently small for n large enough by the rate of convergence of the nonparametric estimator. The second term, $P(\hat{m}(x_i) > 0, m(x_i) \in (-\delta_n, 0))$, is bounded above by $P(m(x_i) \in (-\delta_n, 0))$ which by the smoothness of $m(x_i)$ converges to 0, and hence can be made arbitrarily small. □

To derive the rate of convergence and limiting distribution theory for the feasible estimator where we first estimate choice probability functions and their derivatives nonparametrically, we expand the nonparametric estimators around true functions that are inside the indicator function in Q_{n2}. Then we can follow the approach in Sherman (1994b). Having already established consistency of the estimator, we will first establish root-n consistency and then asymptotic normality. For root-n consistency we will apply Theorem 1 of Sherman (1994b) and so here we change notation to deliberately stay as close as possible to his. We will actually apply this theorem twice, first establishing a slower than root-n consistency result and then root-n consistency. Keeping our notation deliberately as close as possible to Sherman(1994b), here replacing our second stage rank objective function $\hat{Q}_{2,n}(\theta)$ with $\hat{\mathcal{G}}_n(\theta)$, our infeasible objective function $Q^{if}_{n,2}(\theta)$ with $\mathcal{G}_n(\theta)$, and denoting our limiting objective function, previously denoted by $\Gamma_0(\theta)$, by $\mathcal{G}(\theta)$. We have the following theorem:

Theorem B.1. *(From Theorem 1 in Sherman (1994b)).*

If δ_n and ε_n are sequences of positive numbers converging to 0, and

1. $\hat{\theta} - \theta_0 = o_p(\delta_n)$
2. *There exists a neighborhood of θ_0 and a constant $\kappa > 0$ such that $\mathcal{G}(\theta) - \mathcal{G}(\theta_0) \geq \kappa\|\theta - \theta_0\|^2$ for all θ in this neighborhood.*
3. *Uniformly over $O_p(\delta_n)$ neighborhoods of θ_0*

$$\hat{\mathcal{G}}_n(\theta) = \mathcal{G}(\theta) + O_p(\|\theta - \theta_0\|/\sqrt{n}) + o_p(\|\theta - \theta_0\|^2) + O_p(\varepsilon_n)$$

then $\hat{\theta} - \theta_0 = O_p(\max(\varepsilon^{1/2}, n^{-1/2}))$.

Once we use this theorem to establish the rate of convergence of our rank estimator, we can attain limiting distribution theory, which will follow from the following theorem:

Theorem B.2. *(From Theorem 2 in Sherman (1994b)). Suppose $\hat{\theta}$ is $\sqrt{n}$-consistent for θ_0, an interior point of Θ. Suppose also that uniformly over $O_p(n^{-1/2})$ neighborhoods of θ_0,*

$$\hat{\mathcal{G}}_n(\theta) = \frac{1}{2}(\theta - \theta_0)'V(\theta - \theta_0) + \frac{1}{\sqrt{n}}(\theta - \theta_0)'W_n + o_p(1/n) \tag{B.1}$$

where V is a negative definite matrix, and W_n converges in distribution to a $N(0, \Delta)$ random vector. Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, V^{-1}\Delta V^{-1}) \tag{B.2}$$

We first turn attention to applying Theorem B.1 to derive the rate of convergence of our estimator. Having already established consistency of our rank estimator, we turn attention to the second condition in Theorem B.1. To show the second condition, we will first derive an expansion for $\mathcal{G}(\theta)$ around $\mathcal{G}(\theta_0)$. We denote that even though $\mathcal{G}_n(\theta)$ is not differentiable in θ, $\mathcal{G}(\theta)$ is sufficiently smooth for Taylor expansions to apply as the expectation operator is a smoothing operator and the smoothness conditions in Assumptions **RK2**, **RK3**. Taking a second order expansion of $\mathcal{G}(\theta)$ around $\mathcal{G}(\theta_0)$, we obtain

$$\mathcal{G}(\theta) = \mathcal{G}(\theta_0) + \nabla_\beta \mathcal{G}(\theta_0)'(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)'\nabla_{\theta\theta}\mathcal{G}(\theta^*)(\theta - \theta_0) \tag{B.3}$$

where ∇_θ and $\nabla_{\theta\theta}$ denote first and second derivative operators and θ^* denotes an intermediate value. We note that the first two terms of the right hand side of the above equation are 0, the first by how we defined the objective function, and the second by our identification result in Theorem 2.1. Define

$$V \equiv \nabla_{\theta\theta}\mathcal{G}(\theta_0) \tag{B.4}$$

and V is positive definite by Assumption **A3**, so we have

$$(\theta - \theta_0)'\nabla_{\theta\theta}\mathcal{G}(\theta_0)(\theta - \theta_0) > 0 \tag{B.5}$$

$\nabla_{\theta\theta}\mathcal{G}(\theta)$ is also continuous at $\theta = \theta_0$ by Assumptions **RK2** and **RK3**, so there exists a neighborhood of θ_0 such that for all θ in this neighborhood, we have

$$(\theta - \theta_0)'\nabla_{\theta\theta}\mathcal{G}(\theta)(\theta - \theta_0) > 0 \tag{B.6}$$

which suffices for the second condition to hold.

To show the third condition in Theorem B.1, we next establish the form of the remainder term when we replace nonparametric estimators with the true functions they are estimating. Specifically we wish to evaluate the difference between

$$[\mathbf{1}\{\partial_2\hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2\hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \tag{B.7}$$

$$+ \quad \mathbf{1}\{\partial_2\hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2\hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\} \tag{B.8}$$

and

$$[\mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \tag{B.9}$$

$$+ \quad \mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\} \tag{B.10}$$

To establish a representation for this difference, we first simplify notation we write the expressions as:

$$I[\hat{m}_1(\mathbf{x}_i) + \hat{m}_2(\mathbf{x}_j) \geq 0]I[\Delta\mathbf{x}'_{ij}\theta \geq 0] \tag{B.11}$$

$$+ \quad I[\hat{m}_1(\mathbf{x}_i) + \hat{m}_2(\mathbf{x}_j) < 0]I[\Delta\mathbf{x}'_{ij}\theta < 0] \tag{B.12}$$

and

$$I[m_1(\mathbf{x}_i) + m_2(\mathbf{x}_j) \geq 0]I[\Delta\mathbf{x}'_{ij}\theta \geq 0] \tag{B.13}$$

$$+ \quad I[m_1(\mathbf{x}_i) + m_2(\mathbf{x}_j) < 0]I[\Delta\mathbf{x}'_{ij}\theta < 0] \tag{B.14}$$

respectively, where here $\mathbf{x}_i$ denotes the separate components of x_{1i}, x_i, and analogous for $\mathbf{x}_j$. We first explore

$$(I[\hat{m}_1(\mathbf{x}_i) + \hat{m}_2(\mathbf{x}_j) \geq 0] - I[m_1(\mathbf{x}_i) + m_2(\mathbf{x}_j) \geq 0])I[\Delta\mathbf{x}'_{ij}\theta \geq 0]$$

for each i, j inside the double summation:

$$\frac{1}{n(n-1)}\sum_{i\neq j}(I[\hat{m}_1(\mathbf{x}_i) + \hat{m}_2(\mathbf{x}_j) \geq 0] - I[m_1(\mathbf{x}_i) + m_2(\mathbf{x}_j) \geq 0])I[\Delta\mathbf{x}'_{ij}\theta \geq 0] \tag{B.15}$$

An immediate technical difficulty that arises with the above term is the presence of a nonparametric estimator inside the indicator function above. A simple approach to deal with this would be to replace the indicator function with a smoothed indicator function in a fashion analogous to Horowitz (1992), under appropriate conditions on the kernel function and smoothing parameter. Such an approach is not necessary as long as the nonparametric estimator $\hat{m}_1(x_i)$ is asymptotically normal, and asymptotically centered at $m_1(x_i)$, which will be the case with our proposed kernel estimator of the probability function and its derivative. In either approach (smoothed indicator or not) we can show that (B.15) can be represented as:

$$\frac{1}{n(n-1)}\sum_{i\neq j}\phi(0)f_{m_{ij}}(0)\left((\hat{m}_1(\mathbf{x}_i) - m_1(\mathbf{x}_i)) + (\hat{m}_2(\mathbf{x}_j) - m_2(\mathbf{x}_j))\right)I[\Delta\mathbf{x}'_{ij}\theta \geq 0] + o_p(n^{-1}) \tag{B.16}$$

where $\phi(0)$ denotes the standard normal pdf evaluated at 0, $f_{m_{ij}}(0)$ denotes the density function of $m_1(\mathbf{x}_i) + m_2(\mathbf{x}_j)$ evaluated at 0, and the $o_p(n^{-1})$ term is uniform in θ lying in $o_p(1)$ neighborhoods of θ_0. Therefore, uniformly for θ in an $o_p(1)$ neighborhood of θ_0, this remainder term converges to 0 at the rate of convergence of the first stage nonparametric estimator, which under Assumptions

RK3, RK4, RK5, is $o_p(n^{-1/4})$. Thus by repeated application of Theorem B.1, we can conclude that the estimator is root-n consistent. To show that the estimator is also asymptotically normal, we will first derive a linear representation for the term:

$$\frac{1}{n(n-1)}\sum_{i\neq j}\phi(0)f_{m_{ij}}(0)(\hat{m}_1(\mathbf{x}_i)-m_1(\mathbf{x}_i))I[\Delta\mathbf{x}_{ij}'\theta\geq 0] \tag{B.17}$$

As this term is linear in the nonparametric estimator $\hat{m}_1(x_i)$, the desired linear representation follows from arguments used in Khan (2001). One slight difference here compared to Khan (2001) is that here our nonparametric estimators and estimands are each ratios of derivatives. Nonetheless, after linearizing these ratios as done in, e.g. Newey and McFadden (1994). Specifically, we have that B.17 can be expressed as:

$$\frac{1}{n(n-1)}\sum_{i\neq j}\phi(0)f_{m_{ij}}(0)\frac{1}{m_{1den}(\mathbf{x}_i)}(\hat{m}_{1num}(\mathbf{x}_i)-m_{1num}(\mathbf{x}_i))I[\Delta\mathbf{x}_{ij}'\theta\geq 0] \tag{B.18}$$

$$-\frac{1}{n(n-1)}\sum_{i\neq j}\phi(0)f_{m_{ij}}(0)\frac{m_{1num}(\mathbf{x}_i)}{m_{1den}(\mathbf{x}_i)^2}(\hat{m}_{1den}(\mathbf{x}_i)-m_{1den}(\mathbf{x}_i))I[\Delta\mathbf{x}_{ij}'\theta\geq 0] \tag{B.19}$$

where $\hat{m}_{1num}(\mathbf{x}_i)$ denotes the numerator $\{\partial_2\hat{P}^{11}(X_{1,i},X_i)\}$, the estimator of $m_{1num}(\mathbf{x}_i)$ which denotes $\{\partial_2 P^{11}(X_{1,i},X_i)\}$, and $\hat{m}_{1den}(\mathbf{x}_i)$ denotes the denominator $\hat{f}_V(X_i)$, the estimator of $m_{1den}(\mathbf{x}_i)$ which denotes $f_V(X_i)$.

Plugging in the definitions of the kernel estimators of $\hat{m}_{1num}(\mathbf{x}_i)$, and $\hat{m}_{1den}(\mathbf{x}_i)$, results in a third order process. Using arguments in Khan (2001) and Powell, Stock, and Stoker (1989) we can express the third order U process as a second order U process plus an asymptotically negligible remainder term. This is of the form:

$$\frac{1}{n}\sum_{i=1}^{n}\phi(0)\frac{\ell(x_i)}{m_{1den}(\mathbf{x}_i)}(y_{1i}-m_{1num}(\mathbf{x}_i))E\left[I[f_{m_{ij}}(0)\Delta\mathbf{x}_{ij}'\theta\geq 0]|x_i\right] \tag{B.20}$$

where $\ell(x_i)\equiv\frac{-f_X'(x_i)}{f_X(x_i)}$. We note that the function $E\left[f_{m_{ij}}(0)I[\Delta\mathbf{x}_{ij}'\theta\geq 0]|x_i\right]$, which we denote here by $\mathcal{H}(x_i,\theta)$ is a smooth function in θ. We will use this feature to expand $\mathcal{H}(x_i,\theta)$ around $\mathcal{H}(x_i,\theta_0)$. Analogous arguments can be used to attain a linear representation of (B.19), which is of the form:

$$\frac{1}{n}\sum_{i=1}^{n}\phi(0)\frac{\ell_2(x_{1i})m_{1num}(\mathbf{x}_i)}{m_{1den}(\mathbf{x}_i)^2}(y_{2i}-m_{1den}(\mathbf{x}_i))E\left[I[f_{m_{ij}}(0)\Delta\mathbf{x}_{ij}'\theta\geq 0]|x_i\right] \tag{B.21}$$

where $\ell_2(x_{1i})\equiv\frac{-f_{X_1}'(x_{1i})}{f_X(x_{1i})}$. Grouping (B.20) and (B.21) we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi(0)\frac{1}{m_{1den}(\mathbf{x}_i)}\left\{\ell(x_i)(y_{1i}-m_{1num}(\mathbf{x}_i))-\frac{m_{1num}(\mathbf{x}_i)}{m_{1den}(\mathbf{x}_i)}\ell_2(x_{1i})(y_{2i}-m_{1den}(\mathbf{x}_i))\right\}\mathcal{H}(x_i,\theta)$$

(B.22)

Note that by Assumptions **RK2**, **RK3**, $\mathcal{H}(x_i, \theta)$ is smooth in θ implying the expansion

$$\mathcal{H}(x_i, \theta) = \mathcal{H}(x_i, \theta_0) + \nabla_\theta \mathcal{H}(x_i, \theta_0)'(\theta - \theta_0)$$

Thus we can express (B.22) as the which we note is a mean 0 sum

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{1rnki}(\theta - \theta_0) \tag{B.23}$$

where

$$\psi_{1rnki} = \phi(0)\frac{1}{m_{1den}(\mathbf{x}_i)}\left\{\ell(x_i)(y_{1i} - m_{1num}(\mathbf{x}_i)) - \frac{m_{1num}(\mathbf{x}_i)}{m_{1den}(\mathbf{x}_i)}\ell_2(x_{1i})(y_{2i} - m_{1den}(\mathbf{x}_i))\right\}\nabla_\theta \mathcal{H}(x_i, \theta_0) \tag{B.24}$$

We can use identical arguments to attain a linear representation for the $U-$ process:

$$\frac{1}{n(n-1)}\sum_{i\neq j} \phi(0) f_{m_{ij}}(0)\left(\hat{m}_2(\mathbf{x}_j) - m_2(\mathbf{x}_j)\right) I[\Delta\mathbf{x}_{ij}'\theta \geq 0] \tag{B.25}$$

where $\hat{m}_2(\mathbf{x}_j)$ is also a ratio of nonparametric estimators where here the numerator is $\hat{m}_{2n}(\mathbf{x}_j)$ denoting $\{\partial_2 \hat{P}^{10}(X_{1,j}, X_j)\}$, the estimator of $m_{2n}(\mathbf{x}_2)$ which denotes $\{\partial_2 P^{10}(X_{1,j}, X_j)\}$, and $\hat{m}_{2d}(\mathbf{x}_j)$ denotes the denominator $\hat{f}_V(X_j)$, the estimator of $m_{1den}(\mathbf{x}_j)$ which denotes $f_V(X_j)$.

and by using identical arguments it too can be represented as a mean 0 sum denoted here by

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{2rnki} \tag{B.26}$$

where ψ_{2rnki} is defined as:

Finally after grouping the two terms and expanding $\mathcal{H}(x_i, \theta)$ around $\mathcal{H}(x_i, \theta_0)$ we get that (B.16) can be represented as:

$$\frac{1}{n}\sum_{i=1}^{n}(\psi_{1rnki} + \psi_{2rnki})'(\theta - \theta_0) + o_p(n^{-1}) \tag{B.27}$$

Combining our results, from Theorem B.2, we have that

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, V^{-1}\Delta V^{-1}) \tag{B.28}$$

where

$$V = \nabla_{\theta\theta}\mathcal{G}(\theta_0) \tag{B.29}$$

and

$$\Delta = E\left[(\psi_{1rnki} + \psi_{2rnki})(\psi_{1rnki} + \psi_{2rnki})'\right] \tag{B.30}$$

C Nonparametric Factor Structure

Here we describe an estimator for the case where we have a nonparametric factor structure. Recall for this model we had the following relationship between unobservable variables:

$$U = g_0(V) + \bar{\Pi} \tag{C.31}$$

where we assumed that $\bar{\Pi} \perp V$.

Our goal in this more general setup is to identify and estimate both α_0 and g_0. Our identification is based on the condition that

$$x_1 + \alpha_0 - g_0(x) = \tilde{x}_1 - g_0(\tilde{x}).$$

if and only if

$$\partial_2 P^{11}(x_1^{(1)}, x^{(1)})/f_V(x^{(1)}) + \partial_2 P^{10}(\tilde{x}_1^{(1)}, \tilde{x}^{(1)})/f_V(\tilde{x}^{(1)}) = 0.$$

Using the same i, j pair notation as before, this gives gives us, in the nonparametric case,

$$X_{1i} - X_{1j} = \alpha_0 + (g_0(X_i) - g_0(X_j)) \tag{C.32}$$

Note the above equation has a "semi parametric form", loosely related to the model considered in, for example, Robinson (1988). However, we point out crucial differences between what we have above and the standard semi linear model. Here we are trying to identify the intercept α_0 which is usually not identified in the semi linear model as it cannot be separately identified from the nonparametric function. However, note above on the right hand side, we do not just have a nonparametric function of X_i, X_j, but the difference of two *identical* and *additively separable* functions $g_0(\cdot)$. In fact it is this differencing of these functions which enables us to separately identify α_0. Furthermore, as will now see when turning to our estimator of α_0, the structure of the nonparametric component, specifically additive separability of two identical functions of X_i, X_j respectively, can easily be incorporated into our approximation of each of them. From a theoretical perspective separable functions have the advantage of effectively being a one dimensional problem, as there are no interaction terms to have to deal with. It is well known that nonparametric estimation of separable functions do not suffer from the "curse of dimensionality". See, for example Newey (1994).

To motivate our estimator of α_0 in this nonparametric factor structure model, we consider modifying methods used to estimate the semi linear model, which is usually expressed as

$$y_i = x_i'\beta_0 + g(z_i) + \varepsilon_i$$

where y_i denotes the observed dependent variable, x_i, z_i are observed regressors, $g(\cdot)$ is an unknown nuisance function, ε_i is an unobserved disturbance term, and β_0 is the unknown regression coefficient vector which is the parameter of interest. There is a very extensive literature in both econometrics and statistics on estimation and inference methods for this model- see for example Powell (1994) for some references.

One popular way to estimate this model is to use an expansion of basis functions, for example polynomials or splines to approximate $g(\cdot)$, and from a random sample of n observations of (y_i, x_i, z_i) regress y_i on $x_i, b(z_i)$ where $b(z_i)$ denotes the set of basis functions used to approximate $g(\cdot)$. As an illustrative example, assuming z_i were scalar, if one were to use polynomials as basis functions, one would estimate the approximate model,

$$y_i = x_i'\beta_0 + \gamma_1 z_i + +\gamma_2 z_i^2 + \gamma_3 z_i^3 + ...\gamma_{k_n} z_i^{k_n} + u_{in}$$

where k_n is a positive integer smaller than the sample size n, and $\gamma_1, \gamma_2, ...\gamma_{k_n}$ are additional unknown parameters. This has been done by regressing y_i on $x_i, z_i, z_i^2, ...z_i^{k_n}$, and our estimated coefficient of x_i would be the estimator of β_0. The validity of this approach has been shown in, for example, Donald and Newey (1994). Now for our problem at hand, incorporating a nonparametric factor structure, we propose a kernel weighted least squares estimator. The weights are as they were before, assigning great weights to pairs of observations where the sum of derivatives of ratios of choice probabilities are closer to 0.

The dependent variable is identical to as before, the set of n choose 2 pairs $X_{1i} - X_{1j}$. The regressors now reflect the series approximation of $g_0(X_i) - g_0(X_j)$:

$$g_0(X_i) - g_0(X_j) \approx \gamma_1(X_i - X_j) + \gamma_2(X_i^2 - X_j^2) + \gamma_3(X_i^3 - X_j^3) + ...\gamma_{k_n}(X_i^{k_n} - X_j^{k_n})$$

So now our estimator would be to regress $X_{1i} - X_{1j}$ on $1, (X_i - X_j), (X_i^2 - X_j^2), ...(X_i^{k_n} - X_j^{k_n})$, using the same weights $\hat{\omega}_{ij}$ so the estimator of α_0, denoted by $\hat{\alpha}_{NP}$, would be the coefficient on 1. Specifying the asymptotic properties of tis estimator would require additional regularity conditions, notably the rate at which the sequence of integers k_n increases with the sample size n.

We again only outline these regularity conditions here, and only to establish consistency. Since the estimator and proof strategy is very similar to that for the closed form estimator in the online supplement to this paper, here we only state the additional one needed for the nonparametric model in this section.

Assumption BFC (Basis function conditions) The basis function approximation of the unknown factor structure function satisfies the following conditions:

BFC.1 The number of basis functions, k_n, satisfies $k_n \to \infty$ and $k_n/n \to 0$.

BFC.2 For every k_n, the smallest eigenvalue of the matrix

$$E[P_{k_n} P_{k_n}']$$

is bounded away from 0 uniformly in k_n, where

$$P_{k_n} \equiv (1, (X_i - X_j), (X_i - X_j)^2, ...(X_i - X_j)^{k_n})'$$

Theorem C.1. *Under Assumptions* **I**,**K**, **H**, **S**, **PS**, **FK**, **FH**, **BFC**,

$$\hat{\alpha}_{NP} \xrightarrow{p} \alpha_0 \tag{C.33}$$

D Proof of Theorem 4.1

Our first result for this model illustrates how identification can become more difficult. In our first result for this model, we show when $-W$ has a bounded support, say $[a, b]$, then α_0 is not identified if $\alpha_0 > b - a$. To establish this, consider an impostor α such that $\alpha < \alpha_0$. In addition, we consider the case where $\alpha_0 - \alpha + b < \alpha_0 + a$ and $\alpha + b < a + 1$. Such α exists because of the fact that $1 - (b - a) > \alpha_0 > b - a$. Let $\Delta = \alpha_0 - \alpha$ and $(\tilde{W}, \tilde{\eta}_1, \tilde{\eta}_2)$ be mutually independent such that $\tilde{W}$ is distributed as $W - \Delta$, $\tilde{\eta}_2$ is distributed as $\eta_2 - \Delta$, and

$$F_{\tilde{\eta}_1}(e) = \begin{cases} F_{\eta_1}(e) & \text{on } e \leq a, \\ F_{\eta_1}(a) & \text{on } \eta_1 \in (a, a + \Delta], \\ F_{\eta_1}(e - \Delta) & \text{on } e \in (a + \Delta, b + \Delta], \\ \frac{\alpha_0 + a - e}{\alpha_0 + a - b - \Delta} F_{\eta_1}(b) + \frac{e - b - \Delta}{\alpha_0 + a - b - \Delta} F_{\eta_1}(\alpha_0 + a) & \text{on } e \in (b + \Delta, \alpha_0 + a], \\ F_{\eta_1}(e) & \text{on } e \in (\alpha_0 + a, \alpha_0 + b), \\ F_{\eta_1}(\alpha_0 + b) + \frac{e - \alpha_0 - b}{a + 1 + \Delta - \alpha_0 - b}(F_{\eta_1}(a + 1) - F_{\eta_1}(\alpha_0 + b)) & \text{on } e \in (\alpha_0 + b, a + 1 + \Delta], \\ F_{\eta_1}(e - \Delta) & \text{on } e \in (a + \Delta + 1, b + \Delta + 1], \\ F_{\eta_1}(b + 1) + \frac{e - (b + \Delta + 1)}{a + \alpha_0 - b - \Delta}(F_{\eta_1}(a + \alpha_0 + 1) - F_{\eta_1}(b + 1)) & \text{on } e \in (b + \Delta + 1, a + \alpha_0 + 1], \\ F_{\eta_1}(e) & \text{on } e > a + \alpha_0 + 1. \end{cases}$$

Then, because $-\tilde{w} = \Delta - w \in [a + \Delta, b + \Delta]$ and $x_1 = 0, 1$,

$$\begin{aligned} P(Y_1 = 1, Y_2 = 0 | X = x, X_1 = x_1) &= \int F_{\eta_1}(x_1 - w)(1 - F_{\eta_2}(x - w)) f_W(w) dw \\ &= \int F_{\tilde{\eta}_1}(x_1 - \tilde{w})(1 - F_{\tilde{\eta}_2}(x - \tilde{w})) f_{\tilde{w}}(\tilde{w}) d\tilde{w}. \end{aligned}$$

Similarly, because $\alpha - \tilde{w} = \alpha_0 - w \in [\alpha_0 + a, \alpha_0 + b]$ and for $e \in (\alpha_0 + a, \alpha_0 + b] \cup (1 + \alpha_0 + a, 1 + \alpha_0 + b]$, $F_{\tilde{\eta}_1}(e) = F_{\eta_1}(e)$, we have

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1 | X = x, X_1 = x_1) &= \int F_{\eta_1}(x_1 + \alpha_0 - w) F_{\eta_2}(x - w) f_W(w) dw \\ &= \int F_{\eta_1}(x_1 + \alpha - (w + \alpha - \alpha_0)) F_{\eta_2}(x - w) f_W(w) dw \\ &= \int F_{\tilde{\eta}_1}(x_1 + \alpha - \tilde{w}) F_{\tilde{\eta}_2}(x - \tilde{w}) f_{\tilde{w}}(\tilde{w}) d\tilde{w}. \end{aligned}$$

This implies α_0 is not identified from the impostor α.

E Proof of Theorem 4.2

We first show that both σ_0 and the density of η_2 are identified. Note X has full support. This implies the density of V denoted as $f_V(\cdot)$ is identified via

$$f_V(v) = \partial_v E(Y_2|X = v).$$

In addition, we have

$$f_V(\cdot) = f_W * \phi_{\sigma_0}(\cdot),$$

where $*$ denotes the convolution operator. Suppose $f_W(\cdot)$ and σ_0 are not identified so that there exist $f'_W(\cdot)$ and σ' such that

$$f_V(\cdot) = f'_W * \phi_{\sigma'}(\cdot).$$

Without loss of generality, we assume $\sigma' \geq \sigma_0$, otherwise, we can just relabel $f_W(\cdot)$ and $f'_W(\cdot)$. Then we have

$$f_W(\cdot) = f'_W * \phi_{(\sigma'^2 - \sigma_0^2)}.$$

By Assumption B5, we have $\sigma^{'} = \sigma_0$, which implies $f_W(\cdot) = f'_W(\cdot)$.

In the following, we proceed given that $f_W(\cdot)$ and σ_0 are known. Recall $F_{\eta_1}(\cdot)$ and $f_{\eta_2}(\cdot)$ as the CDF and PDF of η_1 and η_2, respectively. Then,

$$P^{11}(x_1, x) = P(Y_1 = 1, Y_2 = 1|X_1 = x_1, X = x) = \int F_{\eta_1}(x_1 + \alpha_0 - \gamma_0 w) F_{\eta_2}(x - \sigma_0 w) f_W(w) dw$$

and

$$P^{10}(x_1, x) = P(Y_1 = 1, Y_2 = 0|X_1 = x_1, X = x) = \int F_{\eta_1}(x_1 - \gamma_0 w)(1 - F_{\eta_2}(x - \sigma_0 w)) f_W(w) dw.$$

Taking derivatives of $P^{11}(x_1, x)$ and $P^{10}(x_1, x)$ w.r.t. x, we have

$$\partial_x P^{11}(x_1, x) = \int F_{\eta_1}(x_1 + \alpha_0 - \gamma_0 w) f_{\eta_2}(x - \sigma_0 w) f_W(w) dw \tag{E.34}$$

and

$$-\partial_x P^{10}(x_1, x) = \int F_{\eta_1}(x_1 - \gamma_0 w) f_{\eta_2}(x - \sigma_0 w) f_W(w) dw. \tag{E.35}$$

Applying Fourier transform on both sides of (E.34) and (E.35), we have

$$\mathcal{F}(\partial_x P^{11}(x_1, \cdot)) = \mathcal{F}_{\sigma_0}(F_{\eta_1}(x_1 + \alpha_0 - \gamma_0 \cdot) f_W(\cdot)) \mathcal{F}(f_{\eta_2}(\cdot)) \tag{E.36}$$

and

$$\mathcal{F}(-\partial_x P^{10}(x_1,\cdot)) = \mathcal{F}_{\sigma_0}(F_{\eta_1}(x_1-\gamma_0\cdot)f_W(\cdot))\mathcal{F}(f_{\eta_2}(\cdot)), \tag{E.37}$$

where for a generic function $g(w)$,

$$\mathcal{F}_{\sigma_0}(g(\cdot))(t) = \frac{1}{\sqrt{2\pi}}\int \exp(-2\pi i t\sigma_0 w)g(w)dw.$$

Then, by (E.36), we can identify $F_{\eta_1}(x_1+\alpha_0-\cdot)$ by

$$F_{\eta_1}(x_1+\alpha_0-\gamma_0\cdot) = \mathcal{F}_{\sigma_0}^{-1}\left(\frac{\mathcal{F}(\partial_x P^{11}(x_1,\cdot))}{\mathcal{F}(f_{\eta_2}(\cdot))}\right)(\cdot)/f_W(\cdot).$$

Similarly, we can identify

$$F_{\eta_1}(x_1-\gamma_0\cdot) = \mathcal{F}_{\sigma_0}^{-1}\left(\frac{\mathcal{F}(-\partial_x P^{10}(x_1,\cdot))}{\mathcal{F}(f_{\eta_2}(\cdot))}\right)(\cdot)/f_W(\cdot),$$

where for a generic function $g(w)$,

$$\mathcal{F}_{\sigma_0}^{-1}(g(\cdot))(t) = \frac{\sigma_0}{\sqrt{2\pi}}\int \exp(2\pi i t\sigma_0 w)g(w)dw.$$

By finding the two pairs $((x_1,w),(x_1',w'))$ and $((\tilde{x}_1,\tilde{w}),(\tilde{x}_1',\tilde{w}'))$ such that $w-w'\neq\tilde{w}-\tilde{w}'$,

$$F_{\eta_1}(x_1+\alpha_0-\gamma_0 w) = F_{\eta_1}(x_1'-\gamma_0 w'), \quad \text{and} \quad F_{\eta_1}(\tilde{x}_1+\alpha_0-\gamma_0\tilde{w}) = F_{\eta_1}(\tilde{x}_1'-\gamma_0\tilde{w}')$$

we can identify both α_0 and γ_0 as the solution of the following linear system:

$$\alpha_0+\gamma_0(w'-w) = x_1'-x_1 \qquad\qquad \alpha_0+\gamma_0(\tilde{w}'-\tilde{w}) = \tilde{x}_1'-\tilde{x}_1.$$

F Partial Identification

In this section, we discuss the information content of factor structure. For illustration purpose, we focus on the "condensed" model:

$$Y_1 = \mathbf{1}\{X_1+\alpha_0 Y_2 - U \geq 0\}$$
$$Y_2 = \mathbf{1}\{X - V \geq 0\}.$$

Assumption 1.

1. $(X_1, X) \perp (U, V)$.

2. (X_1, X) *are continuously distributed with absolute continuous joint density w.r.t. Lebesgue measure. The support of* (X_1, X) *is* $[a,b]\times$ *Supp*(X)*, in which Supp*(X)*, the support of* X*, is compact.*

3. V *is continuously distributed over* $\Re$ *and its density w.r.t. Lebesgue measure exist.*

Theorem F.1. *Assumption 1 holds. (1) Then $|\alpha_0| \leq b-a$ is necessary and sufficient for α_0 to be identified. (2) When $|\alpha_0| > b-a$, the sharp identified set for α_0 is*

$$\mathcal{A}^* = \{\alpha : \alpha > b-a \text{ if } \alpha_0 > 0 \text{ and } \alpha < a-b \text{ if } \alpha_0 < 0\}.$$

Next, we assume, in addition to Assumption 1, the factor structure, i.e., (2.3) in Section 2. Recall in Section 3, under the factor structure, our rank estimator can be written as an M-estimator

$$\hat{\theta} = \arg\max_{\theta} Q_{n,2}(\theta) \equiv \sum_{i \neq j} \hat{g}_{i,j}(\theta)$$

in which

$$\begin{aligned} \hat{g}_{i,j}(\theta) &= [\mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \\ &+ \mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\}], \end{aligned}$$

with

$$\Phi(x_1, x, \tilde{x}_1, \tilde{x}; \theta) = x_1 + \alpha - \gamma x - (\tilde{x}_1 - \gamma \tilde{x}).$$

The information content explored by the M-estimator can be summarized as follows:

$$\begin{aligned} \mathcal{A}_2(\theta) = \{(X_1, \tilde{X}_1, X, \tilde{X}), &\Phi(X_1, X, \tilde{X}_1, \tilde{X}; \theta_0) \geq 0 > \Phi(X_1, X, \tilde{X}_1, \tilde{X}; \theta) \\ &\text{or } \Phi(X_1, X, \tilde{X}_1, \tilde{X}; \theta_0) < 0 \leq \Phi(X_1, X, \tilde{X}_1, \tilde{X}; \theta)\}. \end{aligned}$$

Then we cannot distinguish, from the true parameter θ_0, all impostors in

$$\overline{\mathcal{A}}_2 = \{\theta : P(\mathcal{A}_2(\theta)) = 0\}.$$

In a simple example, if $\text{Supp}(X_1, X) = [a, b] \times [c, d]$, then θ_0 is identified if $|\alpha_0| < b - a + |\gamma_0|(d-c)$. Recall Theorem F.1, without imposing factor structure, the necessary and sufficient condition for achieving identification is $|\alpha_0| \leq b-a$. Therefore, the blue area in the Figure below is the additional parts of parameter space that is identified with factor structure but not otherwise.

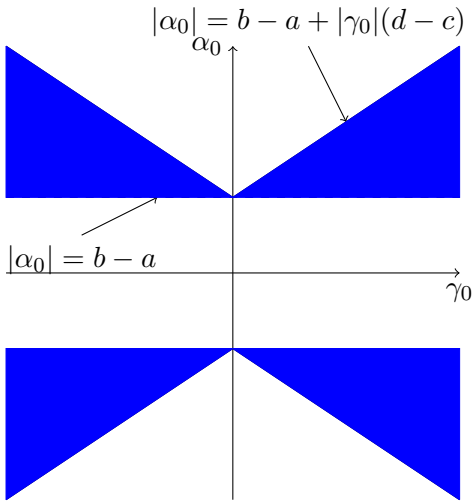

Figure 1: Identifying Power of Factor Structure

When we assume the factor structure, the parameter is still not identified if $|\alpha_0| > b-a+|\gamma_0|(d-c)$. In this case, if we do not impose factor structure, by Theorem F.1(2), the sharp identified set is $\{\alpha : \alpha > b - a\}$ while with the factor structure, the identified set (not necessarily sharp) is $|\alpha| > b - a + |\gamma|(d - c)$. This implies, when identification fails in both cases, the blue area is also the extra identifying power on the identified set given by the factor structure.

G Proof of Theorem F.1

For the first result in the theorem, denote $P^{ij}(x_1, x) = Prob(Y_1 = i, Y_2 = j|X_1 = x_1, X = x)$. Then

$$\begin{aligned} P^{11}(x_1, x) &= \int_{-\infty}^{x} F_U(x_1 + \alpha_0|V = v)f(v)dv \\ P^{10}(\tilde{x}_1, x) &= \int_{x}^{+\infty} F_U(\tilde{x}_1|V = v)f(v)dv. \end{aligned} \tag{G.38}$$

Taking derivatives w.r.t. the second argument of the the LHS function, we have

$$\begin{aligned} \partial_2 P^{11}(x_1, x) &= F_U(x_1 + \alpha_0|V = x)f(x) \\ \partial_2 P^{10}(\tilde{x}_1, x) &= -F_U(\tilde{x}_1|V = x)f(x). \end{aligned}$$

If $|\alpha_0| \leq b - a$, then there exists pair $(x_1, \tilde{x}_1)$ such that $x_1 + \alpha_0 = \tilde{x}_1$. This pair can be identified by checking the equation below:

$$\partial_2 P^{11}(x_1, x)/f(x) + \partial_2 P^{10}(\tilde{x}_1, x)/f(x) = 0.$$

This concludes the sufficient part.

When $\alpha_0 < a - b$, for any $\alpha < \alpha_0$, we can define

$$\begin{aligned} \tilde{U} &= U + \alpha - \alpha_0 && \text{if} && U \leq b + \alpha_0 \\ \tilde{U} &= U && \text{if} && U > b + \alpha_0 \end{aligned}$$

Then for any $x_1 \in [a, b]$,

$$\begin{aligned} P(\tilde{U} \leq x_1 + \alpha | V = v) &= P(\tilde{U} \leq x_1 + \alpha, U \leq b + \alpha_0) + P(\tilde{U} \leq x_1 + \alpha, U > b + \alpha_0 | V = v) \\ &= P(U \leq x_1 + \alpha_0 | V = v) \\ P(\tilde{U} \leq x_1 | V = v) &= P(\tilde{U} \leq x_1, U \leq b + \alpha_0 | V = v) + P(\tilde{U} \leq x_1, U > b + \alpha_0 | V = v) \\ &= P(U \leq b + \alpha_0, U \leq x_1 + \alpha_0 - \alpha | V = v) + P(b + \alpha_0 < U \leq x_1, | V = v) \\ &= P(U \leq b + \alpha_0 | V = v) + P(b + \alpha_0 < U \leq x_1, | V = v) \\ &= P(U \leq x_1 | V = v). \end{aligned}$$

Let $G_{U,V}$ and $G_{\tilde{U},V}$ be the joint distribution of (U, V) and $(\tilde{U}, V)$ respectively. Then the above calculation with (G.38) imply that $(\alpha_0, G_{U,V})$ and $(\alpha, G_{\tilde{U},V})$ are observationally equivalent.

When $\alpha_0 > b - a$, for any $\alpha > \alpha_0$, we can define

$$\begin{aligned} \tilde{U} &= U + \alpha - \alpha_0 && \text{if} && U > a + \alpha_0 \\ \tilde{U} &= U && \text{if} && U \leq a + \alpha_0 \end{aligned}$$

Then for any $x_1 \in [a, b]$,

$$\begin{aligned} P(\tilde{U} \leq x_1 + \alpha | V = v) &= P(\tilde{U} \leq x_1 + \alpha, U \leq a + \alpha_0) + P(\tilde{U} \leq x_1 + \alpha, U > a + \alpha_0 | V = v) \\ &= P(U \leq a + \alpha_0 | V = v) + P(a + \alpha_0 < U \leq x_1 + \alpha_0 | V = v) \\ &= P(U \leq x_1 + \alpha_0 | V = v). \\ P(\tilde{U} \leq x_1 | V = v) &= P(\tilde{U} \leq x_1, U \leq a + \alpha_0 | V = v) + P(\tilde{U} \leq x_1, U > a + \alpha_0 | V = v) \\ &= P(U \leq x_1 | V = v). \end{aligned}$$

So again, $(\alpha_0, G_{U,V})$ and $(\alpha, G_{\tilde{U},V})$ are observationally equivalent.

For the second result in the theorem, first note that, when $|\alpha_0| > b - a$, the sign of α_0 is identified by the data. We take $\alpha_0 > b - a$ as an example. By the proof of Theorem F.1, we have already shown that all $\alpha > \alpha_0$ is in the identified set. Now we consider $\frac{b-a+\alpha_0}{2} \leq \alpha < \alpha_0$.

$$\begin{aligned} \tilde{U} &= U + \alpha - \alpha_0 && \text{if} && U > a + \alpha \\ \tilde{U} &= U && \text{if} && U \leq a + \alpha \end{aligned}$$

Then for any $x_1 \in [a,b]$,

$$\begin{aligned}
P(\tilde{U} \le x_1 + \alpha | V = v) &= P(\tilde{U} \le x_1 + \alpha, U \le a + \alpha) + P(\tilde{U} \le x_1 + \alpha, U > a + \alpha | V = v) \\
&= P(U \le a + \alpha | V = v) + P(a + \alpha < U \le x_1 + \alpha_0 | V = v) \\
&= P(U \le x_1 + \alpha_0 | V = v). \\
P(\tilde{U} \le x_1 | V = v) &= P(\tilde{U} \le x_1, U \le a + \alpha | V = v) + P(\tilde{U} \le x_1, U > a + \alpha | V = v) \\
&= P(U \le x_1 | V = v) + P(U \le x_1 + \alpha_0 - \alpha, U > a + \alpha | V = v). \\
&= P(U \le x_1 | V = v).
\end{aligned}$$

Here note that the last equality is because $x_1 + \alpha_0 - \alpha \le b + \alpha_0 - \alpha \le a + \alpha$ if $\alpha \ge \frac{b-a+\alpha_0}{2}$. Denote $\alpha^{(1)} = \frac{b-a+\alpha_0}{2}$. Then we have shown that there exists $U^{(1)}(\alpha)$ which only depends on α such that for any $x1 \in [a,b]$, any v and any $\alpha_0 > \alpha \ge \alpha^{(1)}$

$$\begin{aligned}
&P(U^{(1)}(\alpha) \le x_1 + \alpha | V = v) = P(U \le x_1 + \alpha_0 | V = v) \\
&P(U^{(1)}(\alpha) \le x_1 | V = v) = P(U \le x_1 | V = v).
\end{aligned}$$

In particular, there exists $U^{(1)}(\alpha^{(1)})$ such that

$$\begin{aligned}
&P(U^{(1)}(\alpha^{(1)}) \le x_1 + \alpha^{(1)} | V = v) = P(U \le x_1 + \alpha_0 | V = v) \\
&P(U^{(1)}(\alpha^{(1)}) \le x_1 | V = v) = P(U \le x_1 | V = v).
\end{aligned}$$

Now repeating the above construction but replacing U with $U^{(1)}$ and α_0 with $\alpha^{(1)}$, we have for any $\alpha^{(1)} > \alpha \ge \alpha^{(2)} \equiv \frac{b-a+\alpha^{(1)}}{2}$, there exists $U^{(2)}(\alpha)$ such that for any $x_1 \in [a,b]$, any v and any $\alpha^{(1)} > \alpha \ge \alpha^{(2)}$,

$$\begin{aligned}
&P(U^{(2)}(\alpha) \le x_1 + \alpha^{(2)} | V = v) = P(U^{(1)}(\alpha^{(1)}) \le x_1 + \alpha^{(1)} | V = v) = P(U \le x_1 + \alpha_0 | V = v) \\
&P(U^{(2)}(\alpha) \le x_1 | V = v) = P(U^{(1)}(\alpha^{(1)}) \le x_1 | V = v) = P(U \le x_1 | V = v).
\end{aligned}$$

This concludes that any α such that $\alpha_0 > \alpha \ge \alpha^{(2)}$ is in the identified set. In general, by repeating the procedure k times, we have that any α such that

$$\alpha_0 > \alpha \ge \alpha^{(k)} = (1 - \frac{1}{2^k})(b - a) + \frac{\alpha_0}{2^k}$$

is in the identified set. For any $\alpha > b-a$, there exists some finite k such that $\alpha > (1-\frac{1}{2^k})(b-a)+\frac{\alpha_0}{2^k}$. This concludes the result that $\alpha > b - a$ is in the identified set.

Finally, since if $\alpha > b-a$, $\partial_2 P^{11}(x_1, x)+\partial_2 P^{10}(\tilde{x}_1, x) > 0$ for all pairs of (x_1, x) and $(\tilde{x}_1, x)$ while, if $\alpha \le b-a$, at least there exists one pair (x_1, x) and $(\tilde{x}_1, x)$ such that $\partial_2 P^{11}(x_1, x)+\partial_2 P^{10}(\tilde{x}_1, x) \le 0$. This implies $\alpha \le b-a$ is not in the identified set. Therefore, the sharp identified set when $\alpha_0 > b-a$ is $\alpha > b - a$.

When $\alpha_0 < a - b$, a symmetric argument implies that the identified set is $\alpha < a - b$.

Online Supplement to "Informational Content of Factor Structures in Simultaneous Binary Response Models": Distribution Theory for Closed Form Estimator

H Distribution Theory for Closed Form Estimator

Many of the basic arguments follow those used in Chen and Khan (2008) and Chen, Khan, and Tang (2016). Recall what the key identification condition that motivated the weighted least squares estimator: For pairs of observations (x_1, x) and $(\tilde{x}_1, \tilde{x})$ in $\text{Supp}(X_1, X)$,

$$x_1 + \alpha_0 - \gamma_0 x = \tilde{x}_1 - \gamma_0 \tilde{x}.$$

if and only if

$$\partial_2 P^{11}(x_1, x)/f_V(x) + \partial_2 P^{10}(\tilde{x}_1, \tilde{x})/f_V(\tilde{x}) = 0.$$

where recall ∂_2 denotes the partial derivative with respect to the second argument. Note that even though the random variable V is unobserved, the density function $f_V(\cdot)$ above can be recovered from the data from the partial derivative of the choice probability in the treatment equation with respect to the regressor in the treatment equation. Thus the above equation involves the sum of two ratios of derivatives of choice probabilities.

Recall $\theta_0 \equiv (\alpha_0, \gamma_0)$. Our estimator of θ_0 is based on pair of observations from the data set. We will denote the random variables of interest with capital letters, for example X_i, X_{1i}, and realizations of them with lower letters, for example x_i, x_{1i}. To denote distinct random variables in the sample when they form pairs, we will use the subscripts i, j.

Note from above, we can express the equation where the pairs receive positive weights (those whose derivatives of choice probabilities summed up to 0) as

$$x_{1i} - x_{1j} = \alpha_0 + \theta_0(x_i - x_j) \tag{H.1}$$

So this motivates regressing the scalar random variable $x_{1i} - x_{1j}$ on the two by one random vector $\mathbf{x}_{ij} \equiv (1, x_i - x_j)$. We can now see that if sufficient such pairs of observations, where the sum of the ratio of derivative of probabilities could be found to equal 0, κ_0 could be recovered as the unique solution to the system of equations corresponding to the pairs, as long as the matrix involving the terms $\mathbf{x}_{ij}$ satisfied a full rank condition. Such an approach is infeasible for two reasons. The first reason is that the probability functions, their derivatives, and hence the ratio of derivatives are unknown. The second reason is that even if these functions were known, if the probability functions are not discrete valued, such "matches" will occur with probability zero.

The first problem can be remedied by replacing the true probability function values with their nonparametric estimates. In the theory here we used a kernel estimator with kernel function

$K(\cdot)$ and bandwidth H_n, whose properties are discussed below. The second problem can be dealt with through the use of "kernel weights" as has been frequently employed in the semiparametric literature.

Specifically, assuming that the ratio of derivatives of conditional probability functions were known, we use the following weighting function for pairs of observations; to illustrate let $P^{k,l,r}$, $k = 0, 1, l = 0, 1$ denotes the ratio of derivatives of choice probabilities. So, for example, $P^{1,1,r} = \partial_2 P^{11}(X_1, X)/f_V(X)$, where ∂_2 denotes the partial derivative with respect to the second argument. Let p_i^{1r}, p_j^{0r} denote the i^{th}, j^{th} realizations of $P^{1,1,r}, P^{1,0,r}$ respectively; then

$$\omega_{ij} = \frac{1}{h_n} k\left(\frac{p_i^{1r} + p_j^{0r}}{h_n}\right) \tag{H.2}$$

In H.2 h_n is a bandwidth sequence, which converges to zero as the sample sizes increases, ensuring that in the limit, only pairs of observations with probability functions summing up to an arbitrarily small number receive positive weight. $k(\cdot)$ is the kernel function, which is symmetric around 0, and assumed to have compact support, integrate to 1, and satisfy certain smoothness conditions discussed later on.

With the weighting matrix defined, a natural estimate of it, $\hat{\omega}_{ij}$ follows from replacing the true probability function values with their nonparametric, e.g. kernel, estimates. This suggests a weighted least squares estimator of $\theta_0 \equiv (\alpha_0, \gamma_0)$, regressing $x_{1i} - x_{1j}$ on $\mathbf{x}_{ij}$, with weights $\hat{\omega}_{ij}$.

Specifically, we propose the following two stage procedure. The first stage is the kernel estimator of the ratio of derivatives of probability functions, and the second stage estimator is defined as:

$$\hat{\theta} = (\sum_{i \neq j} \tau_i \tau_j \hat{\omega}_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}')^{-1} (\sum_{i \neq j} -\tau_i \tau_j \hat{\omega}_{ij} \mathbf{x}_{ij} \Delta x_{1ij}) \tag{H.3}$$

where $\Delta x_{1ij} \equiv x_{1i} - x_{1j}$, $\mathbf{x}_{ij} \equiv (1, x_i - x_j)$ and $\tau_i \equiv \tau(x_{1i}, x_i)$ is a trimming function to remove observations where regressors take values near the boundary of its support.

We will outline the asymptotic properties of this estimator. Here we use similar arguments to this used in Chen and Khan (2008) and keep our notation as close as possible to that used in that paper. To simplify characterizing the asymptotic properties of this estimator and the regularity conditions we impose, we first define the following functions of $P^{k,l,r}$ for $k = l = 1, k = 1, l = 0$ at their i^{th} and j^{th} realized values, denoted by p_i^{1r}, p_j^{0r}

1. $f_{(P_0^{k,l,r}} = f_{P_0^{k,l,r}}(P_{0i}^{k,l,r})$, where $f_{P_0^{k,l,r}}(\cdot)$ denotes the density function of $P_{0i}^{k,l,r}$.
2. $\mu_{\tau i} = E\left[\tau_i | P_{0i}^{k,l,r}\right]$
3. $\mu_{\tau x i} = E\left[\tau_i \tilde{X}_i | P_{0i}^{k,l,r}\right]$

4. $\mu_{\tau xxi} = E\left[\tau_i \tilde{X}_i \tilde{X}_i' | P_{0i}^{k,l,r}\right]$

$\mu_1(p_i^{1r}, p_j^{0r}) \equiv E[\mathbf{x}_{ij}\mathbf{x}_{ij}' | p_i^{1r}, p_j^{0r}]$ where $\mathbf{x}_i$ denotes the 2×1 vector $(1, x_i)$, $\mu_0(p_j^{0r}) \equiv E[\mathbf{x}_j | p_j^{0r}]$, where $\mathbf{x}_j$ denotes the 2×1 vector $(1, x_j)$, $f_1(\cdot)$ denotes the density function of the random variable $P^{1,1,r}$, $f_0(\cdot)$ denotes the density function of the random variable $P^{1,0,r}$.

Our derivation of the asymptotic properties of this estimator are based on the following assumptions[1]:

Assumption I (Identification) The 2×2 matrix:

$$M_1 = E\left[\mu_1(p_i^{1r}, -p_i^{1r})' f_0(-p_i^{1r})\right]$$

has full rank.

Assumption K (Second stage kernel function) The kernel function $k(\cdot)$ used in the second stage (to match the sum of ratios of derivatives to 0) is assumed to have the following properties:

K.1 $k(\cdot)$ is twice continuously differentiable, has compact support and integrates to 1.

K.2 $k(\cdot)$ is symmetric about 0.

K.3 $k(\cdot)$ is an eighth order kernel:

$$\begin{aligned} \int u^l k(u) du &= 0 \quad \text{for } l = 1, 2, 3, 4, 5, 6, 7 \\ \int u^8 k(u) du &\neq 0 \end{aligned}$$

Assumption H (Second stage bandwidth sequence) The bandwidth sequence h_n used in the second stage is of the form:

$$h_n = cn^{-\delta}$$

where c is some constant and $\delta \in (\frac{1}{16}, \frac{1}{12})$.

Assumption S (Order of Smoothness of Density and Conditional Expectation Functions)

S.1 The functions $P^{k,l,r}$ are eighth order continuously differentiable with derivatives that are bounded on the support of τ_i.

S.2 The functions $f_{P_0^{k,l,r}}(\cdot)$ (the density function of the random variable $P^{k,l,r}$) and $E[\mathbf{x}_i | P^{k,l,r} = \cdot]$, where $\mathbf{x}_i$ denotes the 2×1 vector $(1, x_i)$ have order of differentiability of 8, with eight order partial derivatives that are bounded on the support of τ_i.

[1]For notational convenience here we suppress the presence of the trimming function.

The final set of assumptions involve restrictions for the first stage kernel estimator of the ratio of derivatives. This involves smoothness conditions on the choice probabilities $P_{0i}^{k,l,r}$, smoothness and moment conditions on the kernel function, and rate conditions on the first stage bandwidth sequence.

Assumption PS (Order of smoothness of probability functions and regressor density functions) The functions $P^{k,l,r}(\cdot)$ and $f_{X_1,X}(\cdot\cdot)$ (the density function of the random vector (X_1, X)) are continuously differentiable of order p_2, where $p_2 > 5$.

Assumption FK (First stage kernel function conditions) $K(\cdot)$, used to estimate the choice probabilities and their derivatives is an even function, integrating to 1 and is of order p_2 satisfying $p_2 > 5$.

Assumption FH (Rate condition on first stage bandwidth sequence) The first stage bandwidth sequence H_n is of the form:

$$H_n = c_2 n^{-\gamma/k}$$

where c_2 is some constant and γ satisfies:

$$\gamma \in \left(\frac{2}{p_2}\left(\frac{1}{3}+\delta\right), \frac{1}{3}-2\delta\right)$$

where δ is regulated by Assumption **H**.

Theorem H.1. *Let*

$$\psi_i = \psi_{1i} + \psi_{2i} + \psi_{3i} + \psi_{4i} \tag{H.4}$$

where $\psi_{ji} \quad j = 1-4$ *are each mean 0 random variables defined in equations H.15,H.19,H.22,H.24, respectively, then under Assumptions* **I,K,H,S,PS,FK,FH**,

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, M_1^{-1} V_1 M_1^{-1}) \tag{H.5}$$

where

$$V_1 = E[\psi_i \psi_i'] \tag{H.6}$$

Proof: Let $\mathbf{x}_{ij} \equiv (1, (x_i - x_j)), \Delta x_{1ij} \equiv x_{1i} - x_{1j}$. Then we can express:

$$\hat{\theta} - \theta_0 = \left(\frac{1}{n(n-1)} \sum_{i \neq j} \hat{w}_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}'\right)^{-1} \frac{1}{n(n-1)} \sum_{i \neq j} \hat{w}_{ij} \mathbf{x}_{ij} (\Delta x_{1ij} - \mathbf{x}_{ij}' \theta_0)$$

We will first derive a plim for the denominator term and the a linear representation for the numerator. For the denominator term here we aim to establish that the double sum $\frac{1}{n(n-1)}\sum_{i\neq j}\hat{w}_{ij}\mathbf{x}_{ij}\mathbf{x}'_{ij}$ converges in probability to the 2×2 matrix M_1. To do so, note by Assumption K.1 we can expand $\hat{w}_{ij}$ around w_{ij}. The remainder term involves the difference between the nonparametrically estimated derivative functions and the true derivative functions. By Assumptions K, H, S this remainder term is uniformly (over the support of the trimming function $\tau(\cdot)$) $o_p(1)$- see e.g. Henderson, Li, Parmeter, and Yao (2015). It thus suffices to establish the probability limit of $\frac{1}{n(n-1)}\sum_{i\neq j} w_{ij}\mathbf{x}_{ij}\mathbf{x}'_{ij}$. To do so we first wish to determine the functional form of its expectation. For notational ease here we let p_i^{1r}, p_j^{0r} denote i^{th} and j^{th} realized values of $P^{1,1,r}, P^{1,0,r}$ respectively, and $\hat{p}_i^{1r}, \hat{p}_j^{0r}$ denote their nonparametric estimators. Following the same arguments as in Chen and Khan (2008), Chen, Khan, and Tang (2016), we can write the expectation of $w_{ij}\mathbf{x}_{ij}\mathbf{x}'_{ij}$ as

$$\int k((p_i^{1r}+p_j^{0r})/h_n)/h_n\mu_1(p_i^{1r},p_j^{0r}))f_1(p_i^{1r})f_0(p_j^{0r})dp_i^{1r}dp_j^{0r}$$

where $\mu_1(p_i^{1r},p_j^{0r}) \equiv E[\mathbf{x}_{ij}\mathbf{x}'_{ij}|p_i^{1r},p_j^{0r}]$, $f_1(\cdot)$ denotes the density function of the random variable $P^{1,1,r}$, $f_0(\cdot)$ denotes the density function of the random variable $P^{1,0,r}$. Changing variables $u = (p_i^{1r}+p_j^{0r})/h_n$ and taking limits as $h_n \to 0$, yields that the above integral is

$$\int \mu_1(p_i^{1r},-p_i^{1r})f_1(p_i^{1r})f_0(-p_i^{1r})dp_i^{1r} = E\left[\mu_1(p_i^{1r},-p_i^{1r})f_0(-p_i^{1r})\right]$$

which is M_1. We next turn attention to the numerator term. This term is of the form:

$$\frac{1}{n(n-1)}\sum_{i\neq j}\hat{w}_{ij}\mathbf{x}_{ij}(\Delta x_{1ij}-\mathbf{x}'_{ij}\theta_0)$$

Again, we expand $\hat{w}_{ij}$ around w_{ij}. The lead term in this expansion is of the form:

$$\frac{1}{n(n-1)}\sum_{i\neq j} w_{ij}\mathbf{x}_{ij}(\Delta x_{1ij}-\mathbf{x}'_{ij}\theta_0)$$

Note that because $p_i^{1r}+p_j^{0r}=0 \Rightarrow \Delta x_{1ij}=\mathbf{x}'_{ij}\theta_0)$ from our identification result, it follows from Assumptions K,H that the lead term is $o_p(n^{-1/2})$. The linear term in the expansion is of the form

$$\frac{1}{n(n-1)}\sum_{i\neq j} w'_{ij}((\hat{p}_i^{1r}-p_i^{1r})+(\hat{p}_j^{0r}-p_j^{0r}))\mathbf{x}_{ij}(\Delta x_{1ij}-\mathbf{x}'_{ij}\theta_0) \tag{H.7}$$

We will first focus on the term

$$\frac{1}{n(n-1)}\sum_{i\neq j} w'_{ij}(\hat{p}_i^{1r}-p_i^{1r})\mathbf{x}_{ij}(\Delta x_{1ij}-\mathbf{x}'_{ij}\theta_0) \tag{H.8}$$

Recall $\hat{p}_i^{1r}$ denotes a ratio of non parametrically estimated terms and p_i^{1r} denotes the ratio of derivatives. Denote these estimated and true ratios as $\hat{f}_{vi}^{-1}\hat{p}_i^1$, $f_{vi}^{-1}p_i^1$ respectively. Linearizing this ratio, the first term is of the form $f_{vi}^{-1}(\hat{p}_i^1-p_i^1)$. So we wish first to evaluate a representation for

$$\frac{1}{n(n-1)}\sum_{i\neq j} w'_{ij}f_{vi}^{-1}(\hat{p}_i^1-p_i^1)\mathbf{x}_{ij}(\Delta x_{1ij}-\mathbf{x}'_{ij}\theta_0) \tag{H.9}$$

Denoting a kernel estimator of the probability function of the outcome variable as a function of $\vec{x} = (x_1, x)$, by $\hat{p}(\vec{x}) = \frac{\sum_j y_{1j} K_H(\vec{x}_j - \vec{x})}{\sum_j K_H(\vec{x}_j - \vec{x})}$ where $K(\cdot)$ is our kernel function, H our bandwidth, and $K_H(\cdot) \equiv \frac{1}{H} K(\frac{\cdot}{H})$, our estimator of the derivative of the probability function is

$$\hat{p}^1(\vec{x}) = \frac{\sum_k y_{1k} K'_H(\vec{x}_k - \vec{x}) \frac{1}{H} \sum_k K_H(\vec{x}_k - \vec{x}) - \sum_k K'_H(\vec{x}_k - \vec{x}) \frac{1}{H} \sum_k y_{1k} K_H(\vec{x}_k - \vec{x})}{(\sum_k K_H(\vec{x}_k - \vec{x}))^2}$$

We plug in the first of the two terms in the above numerator into H.9 yielding

$$\frac{\frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} w'_{ij} f_{vi}^{-1} (y_{1k} K'_H(\vec{x}_k - \vec{x}_i) \frac{1}{H} - p_i^1) \mathbf{x}_{ij} (\Delta x_{1ij} - \mathbf{x}'_{ij} \theta_0)}{\frac{1}{n} \sum_k K_H(\vec{x}_k - \vec{x}_i)}$$

In the above expression, we replace the denominator term with its plim[2] , which is $f_{\vec{X}}(x_i)$, which gives the expression:

$$\frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} \left(\frac{y_{1k} K'_H(\vec{x}_k - \vec{x}_i) \frac{1}{H}}{f_{\vec{X}}(\vec{x}_i)} - p_i^1 \right) f_{vi}^{-1} \Gamma_{ij} \tag{H.10}$$

where $\Gamma_{ij} = w'_{ij} \mathbf{x}_{ij} (\Delta x_{1ij} - \mathbf{x}'_{ij} \theta_0)$. Evaluating a linear representation for the above third order U statistic in H.10, we first evaluate the expectation of $\frac{1}{f_{\vec{X}}(\vec{x}_i)} y_{1k} K'_H(\vec{x}_k - \vec{x}_i) \frac{1}{H}$ conditioning on $\vec{x}_i$. This can be expressed after a change of variables as

$$\frac{1}{f_{\vec{X}}(\vec{x}_i)} \int p(uH + \vec{x}_i) K'(u) f_{\vec{X}}(uH + \vec{x}_i) du \frac{1}{H}$$

Where here $f_{\vec{X}}(\cdot)$ denotes the density function of $\vec{X}_i$. Next we can expand around $uH = 0$ inside the integral. The lead term is 0 as $K(\cdot)$ vanishes at the boundary of its support. The linear term is $p^1(\vec{x}_i) f_{\vec{X}}(\vec{x}_i) + p(\vec{x}_i) f'_{\vec{X}}(\vec{x}_i)$ using that $\int u K'(u) du = -1$. Thus the conditional expectation of the ratio $\frac{y_{1k} K'_H(\vec{x}_k - \vec{x}_i) \frac{1}{H}}{f_{\vec{X}}(\vec{x}_i)}$ is $p^1(\vec{x}_i) + p(\vec{x}_i) f'_{\vec{X}}(\vec{x}_i) / f_{\vec{X}}(\vec{x}_i)$. The first term, $p^1(\vec{x}_i)$, cancels out with $p^1(\vec{x}_i)$ in H.10. Now, note the second term in H.8, $\frac{\sum_k K'_H(\vec{x}_k - \vec{x}) \frac{1}{H} \sum_k y_{1k} K_H(\vec{x}_k - \vec{x})}{(\sum_k K_H(\vec{x}_k - \vec{x}))^2}$ is by analogous arguments $f'_{\vec{X}}(\vec{x}_i) p(\vec{x}_i) / f_{\vec{X}}(\vec{x}_i) + o_p(n^{-1/2})$. So combining these results one conclusion that can be drawn is an average derivative type result (e.g. Powell, Stock, and Stoker (1989)):

$$\frac{1}{n} \sum_{i=1}^n \hat{p}^1(\vec{x}_i) - p^1(\vec{x}_i) = \frac{1}{n} \sum_{i=1}^n \left\{ y_{1i} \frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^1(\vec{x}_i) \right\} + o_p(n^{-1/2}) \tag{H.11}$$

So plugging H.11 into H.10 yields:

$$\frac{1}{n(n-1)} \sum_{i \neq j} \left\{ y_{1i} \frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^1(\vec{x}_i) \right\} f_{vi}^{-1} \Gamma_{ij} + o_p(n^{-1/2})$$

[2]The resulting remainder term, involving the difference between the denominator term and its plim, can shown to be asymptotically negligible, as shown in Chen, Khan, and Tang (2016)

As an additional step we want a representation for Γ_{ij}. By its definition,

$$\frac{1}{n(n-1)}\sum_{i\neq j}\Gamma_{ij} = \frac{1}{n(n-1)}\sum_{i\neq j} w'_{ij}\mathbf{x}_{ij}(\Delta x_{1ij} - \mathbf{x}'_{ij}\theta_0) = \frac{1}{n(n-1)}\sum_{i\neq j}\frac{1}{h^2}k'\left(\frac{p_i^{1r}+p_j^{0r}}{h}\right)\zeta(\vec{x}_i,\vec{x}_j) \tag{H.12}$$

where $\zeta(\vec{x}_i,\vec{x}_j) \equiv \mathbf{x}_{ij}(\Delta x_{1ij} - \mathbf{x}'_{ij}\theta_0)$. To attain this representation, we evaluate the expectation of the term inside the double summation. We express this as

$$\frac{1}{h^2}\int k'\left(\frac{p_i^{1r}+p_j^{0r}}{h}\right)\bar{\zeta}(p_i^{1r},p_j^{0r})f_1(p_i^{1r})f_0(p_j^{0r})dp_i^{1r}dp_j^{0r}$$

where recall $f_1(\cdot)$ denotes the density function of the random variable $P^{1,1,r}$, $f_0(\cdot)$ denotes the density function of the random variable $P^{1,0,r}$, and here, $\bar{\zeta}(p_i^{1r},p_j^{0r}) \equiv E[\zeta(\vec{x}_i,\vec{x}_j)|p_i^{1r},p_j^{0r}]$ To evaluate the above integral we construct the change of variables $u = \frac{p_i^{1r}+p_j^{r0}}{h}$ and expand inside the integral. Before expanding the integral is of the form

$$\frac{1}{h}\int k'(u)\bar{\zeta}(p_i^{1r},uh-p_i^{1r})f_1(p_i^{1r})f_0(uh-p_i^{1r})dudp_i^{1r}$$

After expanding, the lead term is 0 because the function $k(\cdot)$ vanishes on the boundary of its support. The next term is of the form:

$$\int\left(\bar{\zeta}_2(p_i^{1r},-p_i^{1r})f_1(p_i^{1r})f_0(-p_i^{1r}) + \zeta(p_i^{1r},-p_i^{1r})f_1(p_i^{1r})f_0'(-p_i^{1r})\right)k'(u)ududp_i^{1r}$$

From our identification result the above integral simplifies to $-E[\bar{\zeta}_2(p_i^{1r},-p_i^{1r})f_0(-p_i^{1r})]$ which we will denote by Ξ_1. So plugging this result into H.8 we have the following result:

$$\begin{aligned}\frac{1}{n(n-1)}\sum_{i\neq j} f_{vi}^{-1}(\hat{p}_i^1 - p_i^1)\Gamma_{ij} &= \frac{1}{n}\sum_{i=1}^n \Xi_1 f_{vi}^{-1}\left\{y_{1i}\frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^1(\vec{x}_i)\right\} + o_p(n^{-1/2}) &\text{(H.13)}\\ &\equiv \frac{1}{n}\sum_{i=1}^n \psi_{1i} + o_p(n^{-1/2}) &\text{(H.14)}\end{aligned}$$

where

$$\psi_{1i} = \Xi_1 f_{vi}^{-1}\left\{y_{1i}\frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^1(\vec{x}_i)\right\} \tag{H.15}$$

We next turn attention to the second term in the linearization of the ratio. This is of the form :

$$\frac{1}{n(n-1)}\sum_{i\neq j}\Gamma_{ij}\frac{p_i^1}{f_{vi}^2}(\hat{f}_{vi} - f_{vi}) \tag{H.16}$$

The term $\hat{f}_{vi}$ is our kernel estimator of the derivative of the probability function in the treatment equation: $\hat{f}_{vi} = \frac{\partial}{\partial X_i} E[Y_{2i}|X_i]$. So we can use analogous arguments to attain a linear representation for this U-statistic in H.16 to conclude

$$\frac{1}{n(n-1)} \sum_{i \neq j} \Gamma_{ij} \frac{p_i^1}{f_{vi}^2} (\hat{f}_{vi} - f_{vi}) = \frac{1}{n} \sum_{i=1}^{n} \Xi_1 f_{vi}^{-2} p_i^1 \left\{ y_{2i} \frac{f'_X(x_i)}{f_X(x_i)} - f_V(x_i) \right\} + o_p(n^{-1/2}) \quad \text{(H.17)}$$

$$\equiv \frac{1}{n} \sum_{i=1}^{n} \psi_{2i} + o_p(n^{-1/2}) \quad \text{(H.18)}$$

where

$$\psi_{2i} = \Xi_1 f_{vi}^{-2} p_i^1 \left\{ y_{2i} \frac{f'_X(x_i)}{f_X(x_i)} - f_V(x_i) \right\} \quad \text{(H.19)}$$

Next we can turn attention to the the second term in H.7,

$$\frac{1}{n(n-1)} \sum_{i \neq j} w'_{ij} (\hat{p}_j^{0r} - p_j^{0r}) \mathbf{x}_{ij} (\Delta x_{1ij} - \mathbf{x}'_{ij} \theta_0) \quad \text{(H.20)}$$

The term $\hat{p}_j^{0r} - p^{0r}$ involves the ratio of two derivatives. So we can proceed as before by linearizing this ratio. This will yield the two expressions:

$$\frac{1}{n} \sum_{i=1}^{n} \Xi_1 f_{vi}^{-1} \left\{ y_{1i} \frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^0(\vec{x}_i) \right\} + o_p(n^{-1/2}) \equiv \frac{1}{n} \sum_{i=1}^{n} \psi_{3i} + o_p(n^{-1/2}) \quad \text{(H.21)}$$

where

$$\psi_{3i} = \Xi_1 f_{vi}^{-1} \left\{ y_{1i} \frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^0(\vec{x}_i) \right\} \quad \text{(H.22)}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \Xi_1 f_{vi}^{-2} p_i^0 \left\{ y_{2i} \frac{f'_X(x_i)}{f_X(x_i)} - f_V(x_i) \right\} + o_p(n^{-1/2}) \equiv \frac{1}{n} \sum_{i=1}^{n} \psi_{4i} + o_p(n^{-1/2}) \quad \text{(H.23)}$$

where

$$\psi_{4i} = \Xi_1 f_{vi}^{-2} p_i^0 \left\{ y_{2i} \frac{f'_X(x_i)}{f_X(x_i)} - f_V(x_i) \right\} \quad \text{(H.24)}$$

So collecting all results we can conclude that the estimator has the linear representation:

$$\hat{\theta} - \theta_0 = M_1^{-1} \frac{1}{n} \sum_{i=1}^{n} \psi_i + o_p(n^{-1/2}) \quad \text{(H.25)}$$

where $\psi_i \equiv \psi_{1i} + \psi_{2i} + \psi_{3i} + \psi_{4i}$.

References

Abbring, J., and J. Heckman (2007): "Econometrics Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics, Vol. 6B*, ed. by J. J. Heckman, and E. E. Leamer. North Holland.

Abowd, J. M., and D. Card (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57, 411–445.

Abrevaya, J., J. Hausman, and S. Khan (2010): "Testing for Causal Effects in a Generalized Regression Model with Endogenous Regressors," *Econometrica*, 6, 2043–2061.

Ashworth, J., V. J. Hotz, A. Maurel, and T. Ransom (2017): "Changes Across Cohorts in Wage Returns to Schooling and Early Work Experiences," NBER Working paper #24160.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221.

Bierens, H., and J. Hartog (1988): "Non-Linear Eegression with Discrete Explanatory Variables, with an Application to the Earnings Function," *Journal of Econometrics*, 38(3), 269–299.

Bonhomme, S., and J.-M. Robin (2010): "Generalized Non-Parametric Deconvolution with an Application to Earnings Dynamics," *Review of Economic Studies*, 77, 491–533.

Butucea, C., and C. Matias (2005): "Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model," *Bernoulli*, 11, 309–340.

Carneiro, P., K. Hansen, and J. J. Heckman (2003): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44, 361–422.

Chen, S., and S. Khan (2008): "Rates of Convergence for Estimating Regression Coefficients in Heteroskedastic Discrete Response Models," *Journal of Econometrics*, 117, 245–278.

Chen, S., S. Khan, and X. Tang (2016): "On the Informational Content of Special Regressors in Heteroskedastic Binary Response Models," *Journal of Econometrics*, 193, 162–182.

Chesher, A. (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73(5), 1525–1550.

Chiburis, R. (2010): "Semiparametric Bounds on Treatment Effects," *Journal of Econometrics*, 159(2), 267–275.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

Donald, S., and W. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis*, 50, 30–40.

Geary, R. (1942): "Inherent relations between random variables," *Proceedings of the Royal Irish Academy*, 47, 63–76.

Gourieroux, C., A. Monfort, and J.-P. Renne (2017): "Statistical inference for independent component analysis: Application to structural VAR models," *Journal of Econometrics*, 196, 111–126.

Han, A. (1987): "The The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 303-316.

Han, S., and S. Lee (2018): "Estimation in a Generalization of a Bivariate Probit Models with Dummy Endogenous Regressors," Working Paper.

Han, S., and E. J. Vytlacil (2017): "Identification in a generalization of bivariate probit models with endogenous regressors," *Journal of Econometrics*, 199, 63–73.

Henderson, D., Q. Li, C. Parmeter, and S. Yao (2015): "Gradient-based Smootking Parameter Selection for Nonparametric Regression Estimation," *Journal of Econometrics*, 184, 233–241.

Horowitz, J. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3).

Hyvärinen, A., and E. Oja (2000): *Independent component analysis: algorithms and applications*, vol. 13. Elsevier.

Jochmans, K. (2013): "Pairwise-comparison estimation with nonparametric controls," *Econometrics Journal*, 16, 340–372.

Khan, S. (2001): "Two Stage Rank Estimation of Quantile Index Models," *Journal of Econometrics*, 100, 319–355.

Khan, S., and D. Nekipelov (2018): "Information structure and statistical information in discrete response models," *Quantitative Economics*, 9, 995–1017.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

——— (2018): "Discussion of "Simple Estimators for Invertible Index Models" by Ahn et al.," *Journal of Business & Economic Statistics*, 36, 11–15.

Klein, R., C. Shan, and F. Vella (2015): "Semi-Parametric Estimation of Sample Selection Models with Binary Selection Rules and Binary Outcomes," *Journal of Econometrics*, 185, 82–94.

Klein, R., and R. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Model," *Econometrica*, pp. 387–421.

Lewbel, A. (2000): "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145–177.

Moneta, A., D. E. P. O. Hoyer, and A. Coad (2013): "Causal inference by independent component analysis: Theory and applications," *Oxford Bulletin of Economics and Statistics*, 75, 705–730.

Mourifié, I. (2015): "Sharp Bounds on Treatment Effects in a Binary Triangular System," *Journal of Econometrics*, 187(1), 74–81.

Newey, W. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, pp. 233–253.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden. North Holland.

Powell, J. (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics, Vol. IV.*, ed. by R. F. Engle, and D. McFadden, pp. 2444–2514. North-Holland.

Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, pp. 1403–1430.

Reiersol, O. (1950): "Identifiability of a Linear Relation Between Variables Which are Subject to Error," *Econometrica*, 18, 375–389.

Robinson, P. (1988): "Root-n-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Shaikh, A. M., and E. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables," *Econometrica*, 79(3), 949–955.

Sherman, R. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137.

——— (1994a): "Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators," *Annals of Statistics*, 22, 439–459.

——— (1994b): "U-Processes in the Analysis of a Generalized Semiparametric Regression Estimator," *Econometric Theory*, 10, 372–395.

Tamer, E. (2003): "Incomplete Bivariate Discrete Response Model with Multiple Equilibria," *Review of Economic Studies*, 70, 147–167.

Vuong, Q., and H. Xu (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, pp. 589–610.

Vytlacil, E. J., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757–779.