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.

e ect literature as a way to identify the joint distribution of potential outcomes from the marginals, and then recover the distribution of treatment e ects from this joint distribution.[1] Factor models have been used in a number of di erent 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 identi cation 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 e ect models t into this framework as treatment is typically a binary and endogenous variable in the system, whose e ect 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 re ected 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 identi cation 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. Speci cally, we assume that the unobservables from the treatment equation (V) and the outcome equation (U) are related through the following factor model:

$$U = {}_0 V +$$

where is an unobserved random variable assumed to be distributed independently of V.[2] Our main nding 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 e ect estimation.

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

identi cation in these models without the factor structure. Importantly, our identi cation results are constructive and translate directly into a rank based estimator of the coe cient 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 identi cation results for the parameters of interest in this model. Section 3 then proposes the estimation procedure and establishes its asymptotic properties. Section 4 explores identi cation in more complicated factor structure models which involve nonparametric relationships between unobservables or multiple idiosyncratic

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 misspeci cation in the sense that any estimator of $_0$ based on these conditions will be inconsistent if $(U; V)$ have a di erent bivariate distribution.

The established di culty of identifying $_0$ in semi parametric, i.e., \distribution free" models, and the sensitivity of its identi cation to misspeci cation in parametric models is what motivates the factor structure we add to the above model in this paper. Speci cally, to allow for endogeneity in the form of possible correlation between $U; V$, we augment the model and add the following equation:

$$U = {}_0 V + \tag{2.3}$$

where is an unobserved random variable, assumed to be distributed independently of $(V; Z_1; Z_2; Z_3)$, and $_0$

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 identi cation of the marginals as a whole and directly consider the identi cation 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 $_0$; $_0$ and denote the linear indices by X_1 Z_1^0 $_0$ + Z_3^0 $_0$ and X Z^0 $_0$, where $Z = (Z_1; Z_2)$. In particular, we treat $_0$ as known. In practice, $_0$ can be identi ed and consistently estimated in a rst 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 $_0$ and $_0$ simultaneously with $_0$. Then (2.1) and (2.2) are simpli ed to

$$Y_1 = 1\{X_1 + \ _0 Y_2 \quad U > 0\} \tag{2.4}$$

and

$$Y_2 = 1\{X \quad V > 0\}: \tag{2.5}$$

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

A1 The parameter $_0$ ($_0$; $_0$) is an element of a compact subset of $<^2$.

A2 The vector of unobserved variables, $(U; V;$) 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 ; 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, $P(X \quad X^* = c|X_1 + \ _0 \quad _0 X = X^*_1 \quad _0 X^*) < 1$, where $(X^*; X^*_1)$ are an independent copy of $(X; X_1)$.

A5 Supp$(X_1 + \ _0 \quad _0 X) \setminus$ Supp$(X_1 \quad _0 X) \neq \emptyset$.

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

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

Remark 2.1.

Further assume we can identify $_0$, 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., $@_2 P^{ij}(z_1; z_3; x)$. Then, by the same argument, we can show that

$$@_2 P^{11}(Z_1; Z_3; X) = f_V(X) + @_2 P^{10}(Z_1; Z_3; X) = f_V(X) = 0$$
$$() \quad Z_1^0 \; _0 + Z_3^0 \; _0 + \; _0 \quad _0 X \quad (Z_1^0 \; _0 + Z_3^0 \; _0 \quad _0 X) = 0 :$$

Then, given sufficient variation in X, we can identify $(\ _0; \ _0)$ along with $(\ _0; \ _0)$ 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 $_0; \ _0$. 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 $@_2 P^{ij}(x_1; x)$, 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 $f_V(\ )$. 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

$$@_2 P^{11}(X_1; X) = f_V(X) + @_2 P^{10}(\tilde{X}_1; \tilde{X}) = f_V(\tilde{X}) \quad 0 \quad () \quad X_1 + \quad X \quad (\tilde{X}_1 \quad \tilde{X}) \quad 0:$$

[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 $_0$ 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 $\hat{}$ denote $(\hat{};\hat{})$, our estimator is of the form:

$$\hat{} = \arg\max_{} Q_{n;2}( ) \quad \sum_{i<j} \hat{g}_{i;j}( ) \tag{3.1}$$

in which

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

with

$$(x_1;x;\tilde{x}_1;\tilde{x};) = x_1 + \quad x \quad (\tilde{x}_1 \quad \tilde{x}):$$

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 $_0$ lies in the interior of , 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 $P^{k;l;r}( )$ and $f_{X_1;X}(:)$ (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 ,

$$\sqrt{n}(\hat{\quad}_0) \to N(0; V^{-1} \quad V^{-1}) \tag{3.2}$$

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:

$$\begin{aligned} Y_1 &= 1\mathrm{f}\, X_1 + {}_0 Y_2 \quad U \quad 0\mathrm{g} \\ Y_2 &= 1\mathrm{f}\, X \quad V \quad 0\mathrm{g}; \end{aligned} \tag{4.2}$$

where $U = {}_0 W + {}_1$, $V = W + {}_2$, and $(W; {}_1; {}_2)$ are mutually independent. First we consider the case ${}_0 = 1$ and X_1 is binary, because even in this context, for the baseline case with one idiosyncratic shock, we can identify ${}_0$. But identi cation of ${}_0$ becomes more di cult in this model, as established in the following theorem

Theorem 4.1. *Suppose* (4.2) *holds,* ${}_0$ *is known to be one,* X_1 *is binary, and* W *has a bounded support* $[\ b; \ a]$ *such that* $0{:}5 > b \quad a$ *and* $1 \quad (b \quad a) > {}_0 > b \quad a$*, then* ${}_0$ *is* not *point identi ed.*

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

B1 $U = {}_0 W + {}_1$ and $V = {}_0 W + {}_2$.

B2 W is standard normally distributed.

B3 W, ${}_1$ and ${}_2$ 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

$$f_V \ 2 \ \mathrm{G} = \mathrm{f}g \text{ is a density on } < \text{ s.t. : } g = g^0 \quad \text{for some density } g^0 \text{ implies that } \quad = 0\mathrm{g};$$

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

Assumption B5 e ectively assumes that the distribution of ${}_2$ has tail properties di erent from those of a normal distribution. This type of assumption is made in the deconvolution literature as it is necessary for identi cation 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 identi ed 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 ${}_0$ is known, then Assumption B5 is not necessary.

Theorem 4.2. *If Assumptions* B1{B5 *hold, then* $_0$, $_0$ *and* $_0$ *are identi ed.*

Note that this identi cation result does not require any variation from X_1, which is in spirit close to the one-factor model in our paper and is di erent from the identi cation 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 nite 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

$$Y_1 = 1\mathrm{f}\, X_1 + \ _0 Y_2 \quad U \quad 0\mathrm{g} \tag{5.1}$$

$$Y_2 = 1\mathrm{f}\, X \quad V > 0\mathrm{g} \tag{5.2}$$

where X_1 is binary with success probability 0.6, X has marginal distribution N (0; 1), X_1 and X are mutually independent, $(X_1; X)$? $(V; \ )$, $U = \ _0 V + \ .$ $(V; \ )$ are distributed independently of each other, where V is distributed following a standard normal distribution, and is distributed either standard normal, Laplace, or $T(3)$. The parameters $(\ _0; \ _0) = (\ 0{:}25; 1{:}2)$ or $(0{:}5; 1{:}2)$.

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

$$\mathrm{j} \ \mathrm{j} \quad \text{length of the support of } X;$$

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 (VY051 fn0.9091 -the X

As results from the table indicate, the nite 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 400 as sample sizes doubles, in line with the parametric rate of convergence of our estimator.

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.

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

B Distribution Theory for the Rank Estimator

Recall we de ned our two step rank estimator as follows: Letting $\hat{}$ denote $(\hat{};\hat{})$, our estimator is of the form:

$$\hat{} = \arg\max \hat{Q}_{n;2}( ) \quad \sum_{i \neq j} \hat{g}_{i;j}( )$$

in which

$$\begin{aligned}\hat{g}_{i;j}( ) &= [1\{\hat{@}_2\hat{P}^{11}(X_{1;i}; X_i)/\hat{f}_V(X_i) + \hat{@}_2\hat{P}^{10}(X_{1;j}; X_j)/\hat{f}_V(X_j) \quad 0\}1\{ (X_{1;i}; X_i; X_{1;j}; X_j;) \quad 0\} \\ &+ 1\{\hat{@}_2\hat{P}^{11}(X_{1;i}; X_i)/\hat{f}_V(X_i) + \hat{@}_2\hat{P}^{10}(X_{1;j}; X_j)/\hat{f}_V(X_j) < 0\}1\{ (X_{1;i}; X_i; X_{1;j}; X_j;) < 0\}];\end{aligned}$$

with

$$(x_1; x; x_1; x;) = x_1 + \quad x \quad (x_1 \quad x)$$

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 $Q_{n;2}( )$, in a way analogous to the method introduced in Sherman (1994b). Following Sherman (1994b), we will derive the asymptotic properties of $\hat{}$ in three stages. We will rst 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 $_0$ lies in the interior of , 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 di erentiable.

RK3 (Order of smoothness of probability functions and regressor density functions) The functions $P^{k;l;r}( )$ and $f_{X_{1;i}X}($:-3 j959 Td [(( )]TJ/n8Td [(RK1)]TJ0 g 0 G15 10.9091 Tf 3.03 0 Td [(

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 $\sqrt{n} H_n^{p-2} \to 0$ and $n^{1/4} H_n \to 0$.

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

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

where

$$g_{i,j}(\cdot) = [1\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) \quad 0\} 1\{(X_{1,i}, X_i, X_{1,j}, X_j,) \quad 0\}$$
$$+ \quad 1\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) < 0\} 1\{(X_{1,i}, X_i, X_{1,j}, X_j,) < 0\}];$$

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

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

$$P(\hat{m}(x_i) > 0; m(x_i) < 0) = P(\hat{m}(x_i) > 0; m(x_i) \quad _n) + P(\hat{m}(x_i) > 0; m(x_i) \; 2 \; (\quad _n; 0))$$

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

$$P(\mathrm{j}\hat{m}(x_i) \quad m(x_i)\mathrm{j} > \ _n) \quad P(\mathrm{k}\hat{m}(\ ) \quad m(\ )\mathrm{k} > \ _n)$$

where the notation $\mathrm{k}\hat{m}(\ ) \quad m(\ )\mathrm{k}$ above denotes the sup norm over x_i. The right hand side probability above will be su ciently small for n large enough by the rate of convergence of the nonparametric estimator. The second term, $P(\hat{m}(x_i) > 0; m(x_i) \; 2 \; (\quad _n; 0))$, is bounded above by $P(m(x_i) \; 2 \; (\quad _n; 0))$ 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 rst 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 rst 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, rst 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 $\hat{Q}_{2;n}(\ )$ with $\hat{G}_n(\ )$, our infeasible objective function $Q^{if}_{n;2}(\ )$ with $G_n(\ )$, and denoting our limiting objective function, previously denoted by $_0(\ )$, by $G(\ )$. We have the following theorem:

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

If $_n$ *and* $"_n$ *are sequences of positive numbers converging to 0, and*

1. $\hat{\ } \quad _0 = o_p(\ _n)$
2. *There exists a neighborhood of* $_0$ *and a constant* $\ > 0$ *such that* $G(\ ) \quad G(\ _0) \quad \mathrm{k} \quad _0\mathrm{k}^2$ *for all* $\ $ *in this neighborhood.*
3. *Uniformly over* $O_p(\ _n)$ *neighborhoods of* $_0$

$$\hat{G}_n(\ ) = G(\ ) + O_p(\mathrm{k} \quad _0\mathrm{k}=\sqrt{n}) + o_p(\mathrm{k} \quad _0\mathrm{k}^2) + O_p("_n)$$

then $\hat{\ } \quad _0 = O_p(\max("^{1=2}; n^{\ 1=2}))$.

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 $\hat{}$ is $\sqrt{n}$-consistent for $_0$, an interior point of : Suppose also that uniformly over $O_p(n^{-1=2})$ neighborhoods of $_0$,

$$\hat{G}_n( ) = \frac{1}{2}(\quad _0)'V(\quad _0) + \frac{1}{\sqrt{n}}(\quad _0)'W_n + o_p(1=n) \tag{B.1}$$

where V is a negative de nite matrix, and W_n converges in distribution to a $N(0;)$ random vector. Then

$$\sqrt{n}(\hat{} \quad _0)) \quad N(0; V^{-1} \quad V^{-1}) \tag{B.2}$$

We rst 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 rst derive an expansion for $G( )$ around $G(_0)$. We denote that even though $G_n( )$ is not di erentiable in , $G( )$ is su ciently 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 $G( )$ around $G(_0)$, we obtain

$$G( ) = G(_0) + r \quad G(_0)'(\quad _0) + \frac{1}{2}(\quad _0)'r \quad G( )(\quad _0) \tag{B.3}$$

where r and r denote rst and second derivative operators and denotes an intermediate value. We note that the rst two terms of the right hand side of the above equation are 0, the rst by how we de ned the objective function, and the second by our identi cation result in Theorem 2.1. De ne

$$V \quad r \quad G(_0) \tag{B.4}$$

and V is positive de nite by Assumption A3 , so we have

$$(\quad _0)'r \quad G(_0)(\quad _0) > 0 \tag{B.5}$$

r $G( )$ is also continuous at $= _0$ by Assumptions RK2 and RK3 , so there exists a neighborhood of $_0$ such that for all in this neighborhood, we have

$$(\quad _0)'r \quad G( )(\quad _0) > 0 \tag{B.6}$$

which su ces 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. Speci cally we wish to evaluate the di erence between

$$[1\{\hat{@}P^{11}(X_{1;i}; X_i)=\hat{f}_V(X_i) + \hat{@}P^{10}(X_{1;j}; X_j)=\hat{f}_V(X_j) \quad 0\}1\{ \quad X_{1;i}; X_i; X_{1;j}; X_j;) \quad 0\}] \tag{B.7}$$

$$+ \quad 1\{\hat{@}P^{11}(X_{1;i}; X_i)=\hat{f}_V(X_i) + \hat{@}P^{10}(X_{1;j}; X_j)=\hat{f}_V(X_j) < 0\}1\{ \quad X_{1;i}; X_i; X_{1;j}; X_j;) < 0\} \tag{B.8}$$

and

$$[1f\,@P^{11}(X_{1;i}\,;X_i)=f_V(X_i) + @P^{10}(X_{1;j}\,;X_j)=f_V(X_j) \quad 0g1f \quad (X_{1;i}\,;X_i\,;X_{1;j}\,;X_j\,;\ ) \quad 0g \tag{B.9}$$

$$+ \quad 1f\,@P^{11}(X_{1;i}\,;X_i)=f_V(X_i) + @P^{10}(X_{1;j}\,;X_j)=f_V(X_j) < 0g1f \quad (X_{1;i}\,;X_i\,;X_{1;j}\,;X_j\,;\ ) < 0g \tag{B.10}$$

To establish a representation for this di erence, we rst simplify notation we write the expressions as:

$$I[\hat{m}_1(x_i) + \hat{m}_2(x_j) \quad 0]I[\ x_{ij}^0 \quad 0] \tag{B.11}$$

$$+ \quad I[\hat{m}_1(x_i) + \hat{m}_2(x_j) < 0]I[\ x_{ij}^0 \ < 0] \tag{B.12}$$

and

$$I[m_1(x_i) + m_2(x_j) \quad 0]I[\ x_{ij}^0 \quad 0] \tag{B.13}$$

$$+ \quad I[m_1(x_i) + m_2(x_j) < 0]I[\ x_{ij}^0 \ < 0] \tag{B.14}$$

respectively, where here x_i denotes the separate components of $x_{1i}; x_i$, and analogous for x_j. We rst explore

$$(I[\hat{m}_1(x_i) + \hat{m}_2(x_j) \quad 0] \quad I[m_1(x_i) + m_2(x_j) \quad 0])I[\ x_{ij}^0 \quad 0]$$

for each $i; j$ inside the double summation:

$$\frac{1}{n(n\ Tf\ 8.455\ 0\ Td\ [\ (\ \ )\ n7.272\ 0\ 2tp9091\ Tf\ 41.242\ 0\ Td\ [\ (\ i;\ )\ -167\ (\ j\ )\ ]TJ/F15\ 10.9091\ Tf}$$

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:

$$\frac{1}{n(n-1)}\sum_{i\neq j}(0)f_{m_{ij}}(0)(\hat{m}_1(x_i) - m_1(x_i))I[x_{ij}^0 \quad 0] \tag{B.17}$$

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:

$$\frac{1}{n(n-1)}\sum_{i\neq j}(0)f_{m_{ij}}(0)\frac{1}{m_{1den}(x_i)}(\hat{m}_{1num}(x_i) - m_{1num}(x_i))I[x_{ij}^0 \quad 0] \tag{B.18}$$

$$\frac{1}{n(n-1)}\sum_{i\neq j}(0)f_{m_{ij}}(0)\frac{m_{1num}(x_i)}{m_{1den}(x_i)^2}(\hat{m}_{1den}(x_i) - m_{1den}(x_i))I[x_{ij}^0 \quad 0] \tag{B.19}$$

where $\hat{m}_{1num}(x_i)$ denotes the numerator $\hat{@}_2P^{11}(X_{1;i}; X_i)$, the estimator of $m_{1num}(x_i)$ which denotes $@_2P^{11}(X_{1;i}; X_i)$, and $\hat{m}_{1den}(x_i)$ denotes the denominator $\hat{f}_V(X_i)$, the estimator of $m_{1den}(x_i)$ which denotes $f_V(X_i)$.

Plugging in the definitions of the kernel estimators of $\hat{m}_{1num}(x_i)$, and $\hat{m}_{1den}(x_i)$, 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:

$$\frac{1}{n}\sum_{i=1}^{n}(0)\frac{\grave{}(x_i)}{m_{1den}(x_i)}(y_{1i} - m_{1num}(x_i))E[I[f_{m_{ij}}(0) \quad x_{ij}^0 \quad 0]|x_i \tag{B.20}$$

where $\grave{}(x_i) \quad \frac{f_X^0(x_i)}{f_X(x_i)}$. We note that the function $E\left[f_{m_{ij}}(0)I[x_{ij}^0 \quad 0]|x_i\right]$, which we denote here by $H(x_i;)$ is a smooth function in . We will use this feature to expand $H(x_i;)$ around $H(x_i; _0)$. Analogous arguments can be used to attain a linear representation of (B.19), which is of the form:

$$\frac{1}{n}\sum_{i=1}^{n}(0)\frac{\grave{}_2(x_{1i})m_{1num}(x_i)}{m_{1den}(x_i)^2}(y_{2i} - m_{1den}(x_i))E[I[f_{m_{ij}}(0) \quad x_{ij}^0 \quad 0]|x_i \tag{B.21}$$

where $\grave{}_2(x_{1i})$ that the f

(B.22)

Note that by Assumptions RK2 , RK3 , $H(x_i; \ )$ is smooth in implying the expansion

$$H(x_i; \ ) = H(x_i; \ _0) + r \quad H(x_i; \ _0)^0(\qquad _0)$$

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

$$\frac{1}{n} \sum_{i=1}^{n} \quad _{1rnki} (\qquad _0) \tag{B.23}$$

where

$$_{1rnki} \quad = \quad (0) \frac{1}{m_{1den}(x_i)} \quad `(x_i)(y_{1i} \quad m_{1num}(x_i)) \quad \frac{m_{1num}(x_i)}{m}$$

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:

$$U = g_0(V) + \qquad \text{(C.31)}$$

where we assumed that ? V.

Our goal in this more general setup is to identify and estimate both $_0$ and g_0. Our identi cation is based on the condition that

$$x_1 + \ _0 \quad g_0(x) = \ast_1 \quad g_0(\ast):$$

if and only if

$$@_2P^{11}(x_1^{(1)}; x^{(1)})=f_V(x^{(1)}) + @_2P^{10}(\ast_1^{(1)}; \ast^{(1)})=f_V(\ast^{(1)}) = 0:$$

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

$$X_{1i} \quad X_{1j} = \ _0 + (g_0(X_i) \quad g_0(X_j)) \qquad \text{(C.32)}$$

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 di erences between what we have above and the standard semi linear model. Here we are trying to identify the intercept $_0$ which is usually not identi ed in the semi linear model as it cannot be separately identi ed 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 di erence of two *identical* and *additively separable* functions $g_0(\ )$. In fact it is this di erencing of these functions which enables us to separately identify $_0$. Furthermore, as will now see when turning to our estimator of $_0$, the structure of the nonparametric component, speci cally 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 e ectively being a one dimensional problem, as there are no interaction terms to have to deal with. It is well known that nonparametric

where y_i denotes the observed dependent variable, $x_i ; z_i$ are observed regressors, $g(\ )$ is an unknown nuisance function, $''_i$ is an unobserved disturbance term, and $_0$ is the unknown regression coe cient 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 $g(\ )$, 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 $g(\ )$. As an illustrative example, assuming z_i were scalar, if one were to use polynomials as basis functions, one would estimate the approximate model,

$$y_i = x_i^0 \ _0 + \ _1 z_i + + \ _2 z_i^2 + \ _3 z_i^3 + :::: \ _{k_n} z_i^{k_n} + u_{in}$$

where k_n is a positive integer smaller than the sample size n, and $_1 ; \ _2 ; ::: \ _{k_n}$ 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 coe cient of x_i would be the estimator of $_0$. 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 $X_{1i} \quad X_{1j}$. The regressors now re ect the series approximation of $g_0(X_i) \quad g_0(X_j)$:

$$g_0(X_i) \quad g_0(X_j) \quad \ _1(X_i \quad X_j) + \ _2(X_i^2 \quad X_j^2) + \ _3(X_i^3 \quad X_j^3) + ::: \ _{k_n}(X_i^{k_n} \quad X^{k_n}$$ y2theC.1141967(z)]TJ//Fand

is bounded away from 0 uniformly in k_n, where

$$P_{k_n} \quad (1; (X_i \quad X_j); (X_i \quad X_j)^2; :::(X_i \quad X_j)^{k_n})^0$$

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

$$\hat{}_{NP} \; !^{p} \; _0 \tag{C.33}$$

D Proof of Theorem 4.1

E Proof of Theorem 4.2

We first show that both $_0$ and the density of $_2$ are identified. Note X has full support. This implies the density of V denoted as $f_V(\ )$ is identified via

$$f_V(v) = @_v E(Y_2 \mathrm{j} X = v):$$

In addition, we have

$$f_V(\ ) = f_W \quad _0(\ );$$

where denotes the convolution operator. Suppose $f_W(\ )$ and $_0$ are not identified so that there exist $f_W^0(\ )$ and ^{0} such that

$$f_V(\ ) = f_W^0$$

and

$$F(\ @_x P^{10}(x_1;\ )) = F_{\ 0}(F_{\ 1}(x_1 \quad _0)f_W(\ ))F(\ $$

Theorem F.1. *Assumption 1 holds. (1) Then* $|\beta_0| \geq b - a$ *is necessary and sufficient for* β_0 *to be identified. (2) When* $|\beta_0| > b - a$*, the sharp identified set for* β_0 *is*

$$A = \{\beta : \beta > b - a \text{ if } \beta_0 > 0 \text{ and } \beta < a - b \text{ if } \beta_0 < 0\}.$$

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

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

in which

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

with

$$\beta(x_1, x, \tilde{x}_1, \tilde{x}, \beta) = x_1 + \beta x - (\tilde{x}_1 + \beta \tilde{x}).$$

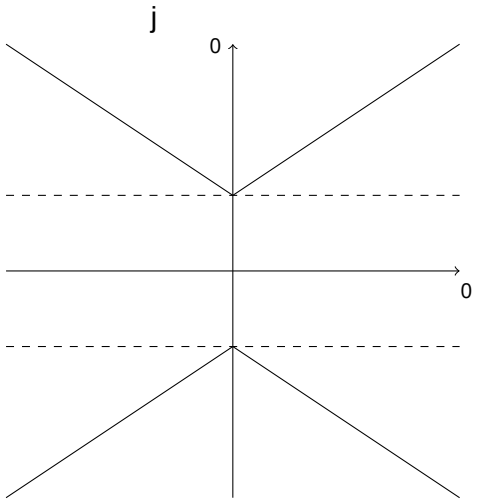
The information content explored by the M-estimator can be summarized as follows:

$$\begin{aligned}A_2(\beta) = \{(X_1, X, \tilde{X}_1, \tilde{X}) : &\beta(X_1, X, \tilde{X}_1, \tilde{X}, \beta_0) \geq 0 > \beta(X_1, X, \tilde{X}_1, \tilde{X}, \beta) \\ &\text{or } \beta(X_1, X, \tilde{X}_1, \tilde{X}, \beta_0) < 0 \leq \beta(X_1, X, \tilde{X}_1, \tilde{X}, \beta)\}.\end{aligned}$$

Then we cannot distinguish, from the true parameter β_0, all impostors in

$$\overline{A}_2 = \{\beta : P(A_2(\beta)) = 0\}.$$

In a simple example, if $\text{Supp}(X_1, X) = [a, b] \times [c, d]$, then β_0 is identified if $|\beta_0| < b - a + |\beta_0|(d - c)$. Recall Theorem F.1, without imposing factor structure, the necessary and sufficient condition for achieving identification is $|\beta_0| \geq b - a$. Therefore, the blue area in the Figure below is the additional parts of parameter space that is identified with factor structure but not otherwise.


j
0
0

When $_0 < a \quad b$, for any $\quad < \ _0$, we can de ne

$$\begin{aligned} \text{Ơ} &= U + \quad _0 && \text{if} && U \quad b + \ _0 \\ \text{Ơ} &= U && \text{if} && U > b + \ _0 \end{aligned}$$

Then for any $x_1 \ 2 \ [a; b]$,

$$\begin{aligned} P(\text{Ơ} \quad x_1 + \ \text{j}V = v) &= P(\text{Ơ} \quad x_1 + \ ; U \quad b + \ _0) + P(\text{Ơ} \quad x_1 + \ ; U > b + \ _0 \text{j}V = v) \\ &= P(U \quad x_1 + \ _0 \text{j}V = v) \\ P(\text{Ơ} \quad x_1 \text{j}V = v) &= P(\text{Ơ} \quad x_1; U \quad b + \ _0 \text{j}V = v) + P(\text{Ơ} \quad x_1; U > b + \ _0 \text{j}V = v) \\ &= P(U \quad b + \ _0; U \quad x_1 + \ _0 \quad \text{j}V = v) + P(b + \ _0 < U \quad x_1; \text{j}V = v) \\ &= P(U \quad b + \ _0 \text{j}V = v) + P(b + \ _0 < U \quad x_1; \text{j}V = v) \\ &= P(U \quad x_1 \text{j}V = v): \end{aligned}$$

Let $G_{U;V}$ and $G_{\text{Ơ};V}$ be the joint distribution of $(U; V)$ and $(\text{Ơ}; V)$ respectively. Then the above calculation with (G.38) imply that $(\ _0; G_{U;V})$ and $(\ ; G_{\text{Ơ};V})$ are observationally equivalent.

When $_0 > b \quad a$, for any $\quad > \ _0$, we can de ne

$$\begin{aligned} \text{Ơ} &= U + \quad _0 && \text{if} && U > a + \ _0 \\ \text{Ơ} &= U && \text{if} && U \quad a + \ _0 \end{aligned}$$

Then for any $x_1 \ 2 \ [a; b]$,

$$\begin{aligned} P(\text{Ơ} \quad x_1 + \ \text{j}V = v) &= P(\text{Ơ} \quad x_1 + \ ; U \quad a + \ _0) + P(\text{Ơ} \quad x_1 + \ ; U > a + \ _0 \text{j}V = v) \\ &= P(U \quad a + \ _0 \text{j}V = v) + P(a + \ _0 < U \quad x_1 + \ _0 \text{j}V = v) \\ &= P(U \quad x_1 + \ _0 \text{j}V = v): \\ P(\text{Ơ} \quad x_1 \text{j}V = v) &= P(\text{Ơ} \quad x_1; U \quad a + \ _0 \text{j}V = v) + P(\text{Ơ} \quad x_1; U > a + \ _0 \text{j}V = v) \\ &= P(U \quad x_1 \text{j}V = v): \end{aligned}$$

So again, $(\ _0; G_{U;V})$ and $(\ ; G_{\text{Ơ};V})$ are observationally equivalent.

For the second result in the theorem , rst note that, when $\text{j} \ _0\text{j} > b \quad a$, the sign of $_0$ is identi ed by the data. We take $_0 > b \quad a$ as an example. By the proof of Theorem F.1, we have already shown that all $\quad > \ _0$ is in the identi ed set. Now we consider $\frac{b \ a+ \ _0}{2} \quad < \ _0$.

$$\begin{aligned} \text{Ơ} &= U + \quad _0 && \text{if} && U > a + \\ \text{Ơ} &= U && \text{if} && U \quad a + \end{aligned}$$

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

$$\begin{aligned} P(U \quad x_1 + \ |V = v) &= P(U \quad x_1 + \ ; U \quad a + \ ) + P(U \quad x_1 + \ ; U > a + \ |V = v) \\ &= P(U \quad a + \ |V = v) + P(a + \ < U \quad x_1 + \ _0|V = v) \\ &= P(U \quad x_1 + \ _0|V = v): \\ P(U \quad x_1|V = v) &= P(U \quad x_1; U \quad a + \ |V = v) + P(U \quad x_1; U > a + \ |V = v) \\ &= P(U \quad x_1|V = v) + P(U \quad x_1 + \ _0 \quad ; U > a + \ |V = v): \\ &= P(U \quad x_1|V = v): \end{aligned}$$

Here note that the last equality is because $x_1 + \ _0 \quad b + \ _0 \quad a +$ if $\frac{b \ a+ \ _0}{2}$. Denote $^{(1)} = \frac{b \ a+ \ _0}{2}$. Then we have shown that there exists $U^{(1)}(\ )$ which only depends on such that for any $x1 \in [a;b]$, any v and any $_0 > \quad ^{(1)}$

$$\begin{aligned} P(U^{(1)}(\ ) \quad x_1 + \ |V = v) &= P(U \quad x_1 + \ _0|V = v) \\ P(U^{(1)}(\ ) \quad x_1|V = v) &= P(U \quad x_1|V = v): \end{aligned}$$

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

$$\begin{aligned} P(U^{(1)}(\ ^{(1)}) \quad x_1 + \ ^{(1)}|V = v) &= P(U \quad x_1 + \ _0|V = v) \\ P(U^{(1)}(\ ^{(1)}) \quad x_1|V = v) &= P(U \quad x_1|V = v): \end{aligned}$$

Now repeating the above construction but replacing U with $U^{(1)}$ and $_0$ with $^{(1)}$, we have for any $^{(1)} > \quad ^{(2)} \quad \frac{b \ a+ \ ^{(1)}}{2}$, there exists $U^{(2)}($

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 identi cation condition that motivated the weighted least squares estimator: For pairs of observations $(x_1; x)$ and $(\tilde{x}_1; \tilde{x})$ in Supp$(X_1; X)$,

$$x_1 + \quad _0 \quad _0 x = \tilde{x}_1 \quad _0 \tilde{x}:$$

if and only if

$$@_2 P^{11}(x_1; x)=f_V(x) + @_2 P^{10}(\tilde{x}_1; \tilde{x})=f_V(\tilde{x}) = 0:$$

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 $f_V(\ )$ 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 $_0 \quad (\ _0; \ _0)$. Our estimator of $_0$

$K(\,)$ 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.

4. $_{xxi} = E\left[{}_i X_i X_i^0 \right] P_{0i}^{k;l;r}$

$_1(p_i^{1r}; p_j^{0r})$ $E[x_{ij}\, x_{ij}^0 \,|\, p_i^{1r}; p_j^{0r}]$ where x_i denotes the 2 1 vector $(1; x_i)$, $_0(p_j^{0r})$ $E[x_j \,|\, p_j^{0r}]$, where x_j denotes the 2 1 vector $(1; x_j)$, $f_1(\,)$ denotes the density function of the random variable $P^{1;1;r}$, $f_0(\,)$ 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 (Identi cation) The 2 2 matrix:

$$M_1 = E \quad {}_1(p_i^{1r};\ p_i^{1r})^0 f_0(\ p_i^{1r})$$

has full rank.

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

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

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

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

$$\int u^l k(u\text{kbml9091 Tf9i]TJ/F15 1d0.9091 Tf 4.242 0 Td 21.8}$$

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 $P^{k;l;r}(\ )$ and $f_{X_1;X}($

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 $\frac{1}{n(n-1)} \sum_{i \neq j} w_{ij}$

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

$$\hat{p}^1(x) = \frac{\sum_k y_{1k} K_H^0(x_k - x) \frac{1}{H} \sum_k K_H(x_k - x) - \sum_k K_H^0(x_k - x) \frac{1}{H} \sum_k y_{1k} K_H(x_k - x)}{(\sum_k K_H(x_k - x))^2}$$

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

$$\frac{\frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} w_{ij}^0 f_{vi}^{-1} (y_{1k} K_H^0(x_k - x_i) \frac{1}{H} - p_i^1) x_{ij} (x_{1ij} - x_{ij}^0{}_0)}{\frac{1}{n} \sum_k K_H(x_k - x_i)}$$

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

$$\frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} \frac{y_{1k} K_H^0(x_k - x_i) \frac{1}{H}}{f_{\bar{X}}(x_i)}$$

As an additional step we want a representation for $_{ij}$. By its de nition,

$$\frac{1}{n(n\quad 1)}\sum_{i\neq j}{}_{ij} = \frac{1}{n(n\quad 1)}\sum_{i\neq j} w_{ij}^{0}\, x_{ij}\,(\quad x_{1ij}\quad x_{ij}^{0}\quad {}_{0}) = \frac{1}{n(n\quad 1)}\sum_{i\neq j}\frac{1}{h^{2}}k^{0}\left(\frac{p_{i}^{1r}+p_{j}^{0r}}{h}\right)(x_{i};x_{j}) \tag{H.12}$$

where $(x_i;x_j)\quad x_{ij}\,(\quad x_{1ij}\quad x_{ij}^{0}\quad {}_{0})$. To attain this representation, we evaluate the expectation of the term inside the double summation. We express this as

$$\frac{1}{h^{2}}\int k^{0}\left(\frac{p_{i}^{1r}+p_{j}^{0r}}{h}\right)(p_{i}^{1r};p_{j}^{0r})\,f_{1}(p_{i}^{1r})\,f_{0}(p_{j}^{0r})\,dp_{i}^{1r}\,dp_{j}^{0r}$$

where recall $f_1(\,)$ denotes the density function of the random variable $P^{1;1;r}$, $f_0(\,)$ denotes the density function of the random variable $P^{1;0;r}$, and here, $(p_i^{1r};p_j^{0r})\quad E[\,(x_i;x_j)|p_i^{1r};p_j^{0r}]$ To evaluate the above integral we construct the change of variables $u = \frac{p_i^{1r}+p_j^{r\,0}}{h}$ and expand inside the integral. Before expanding the integral is of the form

$$\frac{1}{h}\int k^{0}(u)\;(p_{i}^{1r};uh\quad p_{i}^{1r})\,f_{1}(p_{i}^{1r})\,f_{0}(uh\quad p_{i}^{1r})\,du\,dp_{i}^{1r}$$

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

$$\int {}_{2}(p_{i}^{1r};\quad p_{i}^{1r})\,f_{1}(p_{i}^{1r})\,f_{0}(\quad p_{i}^{1r}) + (p_{i}^{1r};\quad p_{i}^{1r})\,f_{1}(p_{i}^{1r})\,f_{0}^{0}(\quad p_{i}^{1r})\;k^{0}(u)\,u\,du\,dp_{i}^{1r}$$

From our identi cation result the above integral simpli es to $E[\,{}_2(p_i^{1r};\quad p_i^{1r})\,f_0(\quad p_i^{1r})]$ which we will denote by ${}_1$. So plugging this result into H.8 we have the following result:

$$\frac{1}{n(n\quad 1)}\sum_{i\neq j} f_{vi}{}^{1}(p_{i}^{1}\quad p_{i}^{1})\;{}_{ij} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\quad {}_{1}f_{vi}{}^{1}\left(\quad y_{1i}\frac{f_{X}^{0}(x_i)}{f_{X}(x_i)}\quad p^{1}(x^{\,})\,f\right.$$

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

$$\frac{1}{n(n-1)} \sum_{i \neq j} \psi_{ij} \frac{p_i^1}{f_{vi}^2} (\hat{f}_{vi} - f_{vi}) = \frac{1}{n} \sum_{i=1}^{n} \psi_{1i} f_{vi}^{-2} p_i^1 \left[y_{2i} \frac{f_X^0(x_i)}{f_X(x_i)} - f_V(x_i) \right] + o_p(n^{-1/2}) \quad \text{(H.17)}$$

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

where

$$\psi_{2i} = \psi_{1i} f_{vi}^{-2} p_i^1 \left[y_{2i} \frac{f_X^0(x_i)}{f_X(x_i)} - f_V(x_i) \right] \quad \text{(H.19)}$$

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

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

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:

$$\frac{1}{n} \sum_{i=1}^{n} \psi_{1i} f_{vi}^{-1} \left[y_{1i} \frac{f_X^0(x_i)}{f_X(x_i)} - p^0(x_i) \right] + o_p(n^{-1/2}) \quad \frac{1}{n} \sum_{i=1}^{n} \psi_{3i} + o_p(n^{-1/2}) \quad \text{(H.21)}$$

where

$$\psi_{3i} = \psi_{1i} f_{vi}^{-1} \left[y_{1i} \frac{f_X^0(x_i)}{f_X(x_i)} - p^0(x_i) \right] \quad \text{(H.22)}$$

and

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

where

$$\psi_{4i} = \psi_{1i} f_{vi}^{-2} p_i^0 \left[y_{2i} \frac{f_X^0(x_i)}{f_X(x_i)} - f_V(x_i) \right] \quad \text{(H.24)}$$

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

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

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 E cient 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,

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.