Supplement to "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

March 2022

Abstract

This paper gathers the supplementary material to the main paper. In Section S.A, we propose an estimator based on our constructive identification strategy and establish its asymptotic properties. Section S.B contains a simulation study. In Section S.C, we establish the asymptotic distribution for the rank estimator. In Section S.D, we consider the identification of the model with two idiosyncratic shocks but no continuous repeated measurements of the common factor. In Sections S.E and S.F, we prove Theorems S.D.1 and S.D.2, respectively. Section S.G discusses the nonparametric factor model and Section S.H establishes the asymptotic properties for the closed-form estimator.

Keywords: Factor Structures, Discrete Choice, Causal Effects.

S.A Estimation and Asymptotic Properties

Our identification result is constructive in the sense that it motivates an estimator for the parameters of interest which we describe in detail here.

As we did in Section C, 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 X of $E[Y_2|X]$, 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_1, X) and $(\tilde{X}_1, \tilde{X})$ is determined by the choice probabilities via

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

This motivates us to use the 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:

$$\hat{\theta} = \arg\max_{\theta} Q_n(\theta), \quad Q_n(\theta) \equiv \sum_{i \neq j} \hat{g}_{i,j}(\theta) \tag{S.A.1}$$

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.[1] 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

[1] An alternative estimation procedure could be based on the exact relationship in (2.7). Note the equality on the left-hand side of (2.7) 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 (S.A.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).

function which is twice continuously differentiable.

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

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.

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 $n^{-1/4}H_n^{-1} \to 0$.

The smoothness condition in Assumption RK4 and Assumption RK5 is due to the fact that we need to nonparametrically estimate $\partial_2 P^{ij}(X_1, X)$ with sufficiently faster convergence rate. This will require a stronger smoothness condition than that required for standard nonparametric estimation. Assumption RK5 ensures that the bias of the first stage estimator of the derivative function converges at the parametric rate and the RMSE of this estimator (with two regressors) is fourth-root consistent, so results for two step estimation in Newey and McFadden (1994) can be applied.

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

Theorem S.A.1. *Under Assumptions* **RK1-RK5**,

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

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

S.B 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

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

$$Y_2 = \mathbf{1}\{X - V > 0\}, \tag{S.B.2}$$

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 Π is distributed following a standard normal distribution, and V 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's (2007) 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's (2007) 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 a 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.

Vytlacil and Yildiz's (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's (2007) estimator is inconsistent in these designs.[2]

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

[2] Because X_1 is binary, Vytlacil and Yildiz's (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.

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 designs we also consider three DGPs (DGPs 1–3) 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 - see Khan and Tamer (2018). In DGP 4, the identification assumptions in both Vytlacil and Yildiz (2007) and our paper hold. In this case, we found that our estimator has similar performance as that proposed by Vytlacil and Yildiz (2007).

The outcome and selection equations are the same as (S.B.1) and (S.B.2), respectively. Then,

DGP 1 : (X_1, X) is jointly standard normally distributed. Let (e_1, e_2) jointly Laplace distributed 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, $U = e_1 + e_2 - 0.5$, 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, support conditions in both Vytlacil and Yildiz (2007) and our paper hold. Table 3 shows that our and Vytlacil and Yildiz's (2007) estimators have similar performance in terms of bias and MSE. Although our estimator is expected to be more efficient as we use the factor structure in estimation, it is not. We conjecture that it is because our estimator does not necessarily use all the information, or in other words, achieve the semiparametric efficiency bound. To establish the semiparametric efficient estimator in the model with and without the factor structure is an interesting yet challenging task. We leave it as a topic for future research.

S.C Proof of Theorem S.A.1

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

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

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

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

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,

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

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 $Q^{if}_{n,2}(\theta)$ 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 S.C.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 S.C.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,*

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

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

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

We first turn attention to applying Theorem S.C.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 S.C.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

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

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

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

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

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

$\nabla_{\theta\theta} \mathcal{G}(\theta)$ 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

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

which suffices for the second condition to hold.

To show the third condition in Theorem S.C.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

$$
\begin{aligned}
& [\mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} && \text{(S.C.7)} \\
+ \quad & \mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\} && \text{(S.C.8)}
\end{aligned}
$$

and

$$
\begin{aligned}
& [\mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} && \text{(S.C.9)} \\
+ \quad & \mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\} && \text{(S.C.10)}
\end{aligned}
$$

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

$$
\begin{aligned}
& I[\hat{m}_1(\mathbf{x}_i) + \hat{m}_2(\mathbf{x}_j) \geq 0] I[\Delta \mathbf{x}_{ij}' \theta \geq 0] && \text{(S.C.11)} \\
+ \quad & I[\hat{m}_1(\mathbf{x}_i) + \hat{m}_2(\mathbf{x}_j) < 0] I[\Delta \mathbf{x}_{ij}' \theta < 0] && \text{(S.C.12)}
\end{aligned}
$$

and

$$
\begin{aligned}
& I[m_1(\mathbf{x}_i) + m_2(\mathbf{x}_j) \geq 0] I[\Delta \mathbf{x}_{ij}' \theta \geq 0] && \text{(S.C.13)} \\
+ \quad & I[m_1(\mathbf{x}_i) + m_2(\mathbf{x}_j) < 0] I[\Delta \mathbf{x}_{ij}' \theta < 0] && \text{(S.C.14)}
\end{aligned}
$$

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:

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

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 (S.C.15) can be represented as:

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

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 S.C.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:

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

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 S.C.17 can be expressed as:

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

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

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:

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

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 (S.C.19), which is of the form:

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

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

$$\frac{1}{n}\sum_{i=1}^{n}\phi(0)\frac{1}{m_{1den}(\mathbf{x}_i)}\left\{\ell(x_i)(y_{1i}-m_{1num}(\mathbf{x}_i))-\frac{m_{1num}(\mathbf{x}_i)}{m_{1den}(\mathbf{x}_i)}\ell_2(x_{1i})(y_{2i}-m_{1den}(\mathbf{x}_i))\right\}\mathcal{H}(x_i,\theta) \tag{S.C.22}$$

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 (S.C.22) as the which we note is a mean 0 sum

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

where

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

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

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

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

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

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 (S.C.16) can be represented as:

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

Combining our results, from Theorem S.C.2, we have that

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

where

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

and

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

S.D Model with Two Idiosyncratic Shocks

In this section, we focus on the identification of (α_0,γ_0) in the "condensed" model that $X_1=Z_1'\lambda_0+Z_3'\beta_0$ is observed and

$$\begin{aligned} Y_1&=\mathbf{1}\{X_1+\alpha_0Y_2-U\geq 0\}\\ Y_2&=\mathbf{1}\{X-V\geq 0\}. \end{aligned} \tag{S.D.31}$$

with the understanding that (λ_0,β_0) can be identified jointly with α_0 and γ_0, as shown in Theorems 2.1 and 3.1. We further impose $U=\gamma_0W+\eta_1$, $V=W+\eta_2$, 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 without the help of repeated measurements, as established in the following theorem.

Theorem S.D.1. *Suppose (S.D.31) 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

C1 $U = \gamma_0 W + \eta_1 \quad \text{and} \quad V = \sigma_0 W + \eta_2.$

C2 W is standard normally distributed.

C3 W, η_1 and η_2 are mutually independent.

C4 X has full support.

C5 Denote the density of η_2 as f_{η_2}, then f_{η_2} does not have a Gaussian component in the sense that

$$f_{\eta_2} \in \mathcal{G} = \{g \text{ is a density on } \Re \text{ s.t. : } g = g' * \phi_\sigma \text{ for some density } g' \text{ implies that } \sigma = 0\},$$

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

Assumption **C5** 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).[3] 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. In our case, note $V = \sigma_0 W + \eta_2$ where W is standard normal and the density of V is identified from data. Here we want to identify σ_0 and the density of η_2. If η_2 has a Gaussian component, then

$$\eta_2 = \eta_2' + \tilde{\sigma}\tilde{W},$$

where $\tilde{W}$ is a standard normal random variable that is independent of η_2' and W and $\tilde{\sigma} > 0$. It implies

$$V = (\sigma_0 W + \tilde{\sigma}\tilde{W}) + \eta_2',$$

where η_2' does not have a Gaussian component. In addition, note that $(\sigma_0 W + \tilde{\sigma}\tilde{W}) = \sqrt{\sigma_0^2 + \tilde{\sigma}^2}G$, for some standard normal random variable G. Therefore, without Assumption **B5**, σ_0 is not identified.

[3] 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 **C5** is not necessary.

Theorem S.D.2. *If Assumptions **C1**–**C5** 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.

S.E Proof of Theorem S.D.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

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

$$= \int F_{\eta_1}(x_1 + \alpha - (w + \alpha - \alpha_0))F_{\eta_2}(x - w)f_W(w)dw$$
$$= \int F_{\tilde{\eta}_1}(x_1 + \alpha - \tilde{w})F_{\tilde{\eta}_2}(x - \tilde{w})f_{\tilde{w}}(\tilde{w})d\tilde{w}.$$

This implies α_0 is not identified from the impostor α.

S.F Proof of Theorem S.D.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

$$f_V(\cdot) = f_{\eta_2} * \phi_{\sigma_0}(\cdot),$$

where $*$ denotes the convolution operator. Suppose $f_{\eta_2}(\cdot)$ and σ_0 are not identified so that there exist $f'_{\eta_2}(\cdot)$ and σ' such that

$$f_V(\cdot) = f'_{\eta_2} * \phi_{\sigma'}(\cdot).$$

Without loss of generality, we assume $\sigma' \geq \sigma_0$, otherwise, we can just relabel $f_{\eta_2}(\cdot)$ and $f'_{\eta_2}(\cdot)$. Then we have

$$f_{\eta_2}(\cdot) = f'_{\eta_2} * \phi_{(\sigma'^2 - \sigma_0^2)}.$$

By Assumption **B5**, we have $\sigma^{'} = \sigma_0$, which implies $f_{\eta_2}(\cdot) = f'_{\eta_2}(\cdot)$.

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

$$\partial_x P^{11}(x_1, x) = \int F_{\eta_1}(x_1 + \alpha_0 - \gamma_0 w) f_{\eta_2}(x - \sigma_0 w) f_W(w) dw \tag{S.F.32}$$

and

$$-\partial_x P^{10}(x_1, x) = \int F_{\eta_1}(x_1 - \gamma_0 w) f_{\eta_2}(x - \sigma_0 w) f_W(w) dw. \tag{S.F.33}$$

Applying Fourier transform on both sides of (S.F.32) and (S.F.33), we have

$$\mathcal{F}(\partial_x P^{11}(x_1, \cdot)) = \mathcal{F}_{\sigma_0}(F_{\eta_1}(x_1 + \alpha_0 - \gamma_0 \cdot) f_W(\cdot)) \mathcal{F}(f_{\eta_2}(\cdot)) \tag{S.F.34}$$

and

$$\mathcal{F}(-\partial_x P^{10}(x_1, \cdot)) = \mathcal{F}_{\sigma_0}(F_{\eta_1}(x_1 - \gamma_0 \cdot) f_W(\cdot)) \mathcal{F}(f_{\eta_2}(\cdot)), \tag{S.F.35}$$

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 (S.F.34), 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:

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

S.G Nonparametric Factor Structure

In this section 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:

$$U = g_0(V) + \tilde{\Pi} \tag{S.G.36}$$

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

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

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

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

$$X_{1i} - X_{1j} = \alpha_0 + (g_0(X_i) - g_0(X_j)) \tag{S.G.37}$$

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 this 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

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

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

$$\hat{\alpha}_{NP} \xrightarrow{p} \alpha_0 \tag{S.G.38}$$

S.H Distribution Theory for Closed Form Estimator

Many of the basic arguments follow those used in Chen and Khan (2008) and Chen et al. (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 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

$$x_{1i} - x_{1j} = \alpha_0 + \theta_0(x_i - x_j) \tag{S.H.39}$$

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, θ_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

$$\omega_{ij} = \frac{1}{h_n} k\left(\frac{p_i^{1r} + p_j^{0r}}{h_n}\right) \tag{S.H.40}$$

In (S.H.40) 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:

$$\hat{\theta} = (\sum_{i \neq j} \tau_i \tau_j \hat{\omega}_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}')^{-1} (\sum_{i \neq j} -\tau_i \tau_j \hat{\omega}_{ij} \mathbf{x}_{ij} \Delta x_{1ij}) \tag{S.H.41}$$

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

$\mu_1(p_i^{1r}, p_j^{0r}) \equiv E[\mathbf{x}_{ij}\mathbf{x}_{ij}' | p_i^{1r}, p_j^{0r}]$ 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[4]:

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:

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

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

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.

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 S.H.1. *Let*

$$\psi_i = \psi_{1i} + \psi_{2i} + \psi_{3i} + \psi_{4i} \tag{S.H.42}$$

where $\psi_{ji} \quad j = 1{-}4$ *are each mean 0 random variables defined in equations S.H.53,S.H.57,S.H.60,S.H.62, respectively, then under Assumptions* **I,K,H,S,PS,FK,FH**,

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, M_1^{-1} V_1 M_1^{-1}) \tag{S.H.43}$$

where

$$V_1 = E[\psi_i \psi_i'] \tag{S.H.44}$$

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 et al. (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 et al. (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

$$\frac{1}{n(n-1)} \sum_{i \neq j} w'_{ij} ((\hat{p}_i^{1r} - p_i^{1r}) + (\hat{p}_j^{0r} - p_j^{0r})) \mathbf{x}_{ij} (\Delta x_{1ij} - \mathbf{x}'_{ij} \theta_0) \tag{S.H.45}$$

We will first focus on the term

$$\frac{1}{n(n-1)} \sum_{i \neq j} w'_{ij} (\hat{p}_i^{1r} - p_i^{1r}) \mathbf{x}_{ij} (\Delta x_{1ij} - \mathbf{x}'_{ij} \theta_0) \tag{S.H.46}$$

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

$$\frac{1}{n(n-1)} \sum_{i \neq j} w'_{ij} f_{vi}^{-1} (\hat{p}_i^1 - p_i^1) \mathbf{x}_{ij} (\Delta x_{1ij} - \mathbf{x}'_{ij} \theta_0) \tag{S.H.47}$$

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 S.H.47 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[5] , which is $f_{\vec{X}}(x_i)$, which

[5]The resulting remainder term, involving the difference between the denominator term and its plim, can shown to be asymptotically negligible, as shown in Chen et al. (2016)

gives the expression:

$$\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k}\left(\frac{y_{1k}K'_H(\vec{x}_k-\vec{x}_i)\frac{1}{H}}{f_{\vec{X}}(\vec{x}_i)}-p_i^1\right)f_{vi}^{-1}\Gamma_{ij} \tag{S.H.48}$$

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 S.H.48, 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 $\int uK'(u)du=-1$. 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 S.H.48. Now, note the second term in S.H.46, $\frac{\sum_k K'_H(\vec{x}_k-\vec{x})\frac{1}{H}\sum_k y_{1k}K_H(\vec{x}_k-\vec{x})}{(\sum_k K_H(\vec{x}_k-\vec{x}))^2}$ 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 et al. (1989)):

$$\frac{1}{n}\sum_{i=1}^{n}\hat{p}^1(\vec{x}_i)-p^1(\vec{x}_i)=\frac{1}{n}\sum_{i=1}^{n}\left\{y_{1i}\frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)}-p^1(\vec{x}_i)\right\}+o_p(n^{-1/2}) \tag{S.H.49}$$

So plugging S.H.49 into S.H.48 yields:

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

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

$$\frac{1}{n(n-1)}\sum_{i\neq j}\Gamma_{ij}=\frac{1}{n(n-1)}\sum_{i\neq j}w'_{ij}\mathbf{x}_{ij}(\Delta x_{1ij}-\mathbf{x}'_{ij}\theta_0)=\frac{1}{n(n-1)}\sum_{i\neq j}\frac{1}{h^2}k'\left(\frac{p_i^{1r}+p_j^{0r}}{h}\right)\zeta(\vec{x}_i,\vec{x}_j) \tag{S.H.50}$$

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 S.H.46 we have the following result:

$$\frac{1}{n(n-1)}\sum_{i\neq j} f_{vi}^{-1}(\hat{p}_i^1 - p_i^1)\Gamma_{ij} = \frac{1}{n}\sum_{i=1}^{n}\Xi_1 f_{vi}^{-1}\left\{y_{1i}\frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^1(\vec{x}_i)\right\} + o_p(n^{-1/2}) \quad \text{(S.H.51)}$$

$$\equiv \frac{1}{n}\sum_{i=1}^{n}\psi_{1i} + o_p(n^{-1/2}) \quad \text{(S.H.52)}$$

where

$$\psi_{1i} = \Xi_1 f_{vi}^{-1}\left\{y_{1i}\frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^1(\vec{x}_i)\right\} \quad \text{(S.H.53)}$$

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

$$\frac{1}{n(n-1)}\sum_{i\neq j}\Gamma_{ij}\frac{p_i^1}{f_{vi}^2}(\hat{f}_{vi} - f_{vi}) \quad \text{(S.H.54)}$$

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 (S.H.54) to conclude

$$\frac{1}{n(n-1)}\sum_{i\neq j}\Gamma_{ij}\frac{p_i^1}{f_{vi}^2}(\hat{f}_{vi} - f_{vi}) = \frac{1}{n}\sum_{i=1}^{n}\Xi_1 f_{vi}^{-2}p_i^1\left\{y_{2i}\frac{f'_X(x_i)}{f_X(x_i)} - f_V(x_i)\right\} + o_p(n^{-1/2}) \quad \text{(S.H.55)}$$

$$\equiv \frac{1}{n}\sum_{i=1}^{n}\psi_{2i} + o_p(n^{-1/2}) \quad \text{(S.H.56)}$$

where

$$\psi_{2i} = \Xi_1 f_{vi}^{-2} p_i^1 \left\{ y_{2i} \frac{f'_X(x_i)}{f_X(x_i)} - f_V(x_i) \right\} \tag{S.H.57}$$

Next we can turn attention to the the second term in (S.H.45),

$$\frac{1}{n(n-1)} \sum_{i \neq j} w'_{ij} (\hat{p}_j^{0r} - p_j^{0r}) \mathbf{x}_{ij} (\Delta x_{1ij} - \mathbf{x}'_{ij} \theta_0) \tag{S.H.58}$$

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:

$$\frac{1}{n} \sum_{i=1}^{n} \Xi_1 f_{vi}^{-1} \left\{ y_{1i} \frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^0(\vec{x}_i) \right\} + o_p(n^{-1/2}) \equiv \frac{1}{n} \sum_{i=1}^{n} \psi_{3i} + o_p(n^{-1/2}) \tag{S.H.59}$$

where

$$\psi_{3i} = \Xi_1 f_{vi}^{-1} \left\{ y_{1i} \frac{f'_{\vec{X}}(\vec{x}_i)}{f_{\vec{X}}(\vec{x}_i)} - p^0(\vec{x}_i) \right\} \tag{S.H.60}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \Xi_1 f_{vi}^{-2} p_i^0 \left\{ y_{2i} \frac{f'_X(x_i)}{f_X(x_i)} - f_V(x_i) \right\} + o_p(n^{-1/2}) \equiv \frac{1}{n} \sum_{i=1}^{n} \psi_{4i} + o_p(n^{-1/2}) \tag{S.H.61}$$

where

$$\psi_{4i} = \Xi_1 f_{vi}^{-2} p_i^0 \left\{ y_{2i} \frac{f'_X(x_i)}{f_X(x_i)} - f_V(x_i) \right\} \tag{S.H.62}$$

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

$$\hat{\theta} - \theta_0 = M_1^{-1} \frac{1}{n} \sum_{i=1}^{n} \psi_i + o_p(n^{-1/2}) \tag{S.H.63}$$

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

References

Abrevaya, J., J. Hausman, and S. Khan (2010). Testing for causal effects in a generalized regression model with endogenous regressors. *Econometrica 78*(6), 2043–2061.

Butucea, C. and C. Matias (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli 11*(2), 309–340.

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.

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.

Han, A. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics 35*(2–3), 303–316.

Henderson, D., Q. Li, C. Parmeter, and S. Yao (2015). Gradient-based smoothing 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).

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 E. Tamer (2018). Discussion of "simple estimators for invertible index models" by ahn et al. *Journal of Business & Economic Statistics 36*, 11–15.

Newey, W. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 233–253.

Newey, W. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics, Vol. 4*. North Holland.

Powell, J. (1994). Estimation of semiparametric models. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics, Vol. IV.*, pp. 2444–2514. North-Holland.

Powell, J., J. Stock, and T. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica*, 1403–1430.

Reiersol, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica 18*(4), 375–389.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Sherman, R. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica 61*, 123–137.

Sherman, R. (1994a). Maximal inequalities for degenerate u-processes with applications to optimization estimators. *Annals of Statistics 22*, 439–459.

Sherman, R. (1994b). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric Theory 10*, 372–395.

Vytlacil, E. J. and N. Yildiz (2007). Dummy endogenous variables in weakly separable models. *Econometrica 75*(3), 757–779.