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

December 2020

**Abstract**

This paper gathers the supplementary material to the original paper. In Section A, we discuss the identification power of the factor structure. In Section B, we propose an estimator based on our constructive identification strategy and establish its asymptotic properties. Section C contains a simulation study. In Sections D and E, we prove Theorems 2.1 and 3.1, respectively. In Section F, we establish the asymptotic distribution for the rank estimator. In Section G, we consider the identification of the model with two idiosyncratic shocks but no continuous repeated measurements of the common factor. In Sections H, I, J and K, we prove Theorems A.1, A.2, G.1 and G.2, respectively.

**Keywords:** Factor Structures, Discrete Choice, Causal Effects.

# A Identification with and without Factor Structure

## A.1 Identification Without Auxiliary Measurements

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

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

**Assumption 1.**

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

*2.* $(X_1, X)$ *are continuously distributed with absolute continuous joint density w.r.t. Lebesgue measure. The conditional support of* $X_1$ *given* $X$ *is* $[a, b]$.

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

**Theorem A.1.** *If Assumption 1 holds, then $|\alpha_0| \leq b - a$ is necessary and sufficient for $\alpha_0$ to be identified.*

We note that under Assumption 1, $|\alpha_0| \leq b - a$ is equivalent to the fact that we can find $x_1$ and $\tilde{x}_1$ in the support of $X_1$ such that $\alpha_0 = x_1 - \tilde{x}_1$.

Next, we assume, in addition to Assumption 1, the factor structure, i.e., (2.6) in Section 2. Our rank estimator can be written as an M-estimator

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

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 will study the asymptotic properties of this estimator in Section B.

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

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

Then we cannot distinguish, from the true parameter $\theta_0$, all impostors in

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

In the condensed model, if $\text{Supp}(X_1, X) = [a, b] \times [c, d]$, then $\theta_0$ is identified if $|\alpha_0| < b - a + |\gamma_0|(d - c)$. Recall Theorem A.1, without imposing factor structure, the necessary and sufficient condition for achieving identification is $|\alpha_0| \leq b - a$. Therefore, the blue area in the Figure below is the additional parts of parameter space that are identified with factor structure but not otherwise.

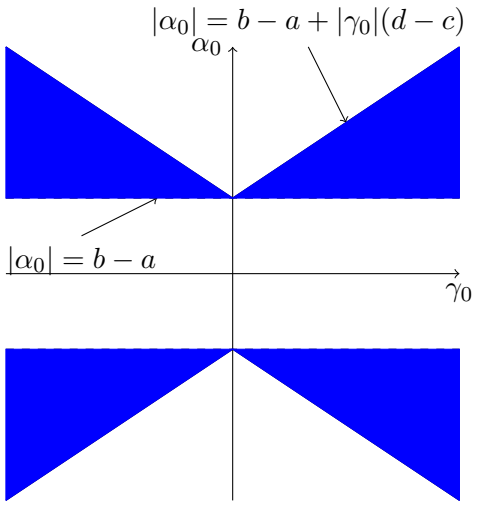


Figure 1: Identifying Power of Factor Structure

**Theorem A.2.** *Assumption 1 holds. When $|\alpha_0| > b - a$, the sharp identified set for $\alpha_0$ is*

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

Theorem A.2 highlights that, in the case without the factor structure and $\alpha_0$ does not satisfy the parameter restriction, except for the fact that the sign of $\alpha_0$ is identified, we actually cannot say much about the value of $|\alpha_0|$. When we assume the factor structure, the parameter is still not identified if $|\alpha_0| > b - a + |\gamma_0|(d - c)$. In addition, suppose $\alpha_0 > 0$. In this case, if we do not impose factor structure, by Theorem A.2, the sharp identified set is $\{\alpha : \alpha > b - a\}$ while with the factor structure, the identified set (not necessarily sharp) is $\alpha > b - a + |\gamma|(d - c)$. This implies, when identification fails in both cases, the blue area is also the extra identifying power on the identified set given by the factor structure.

## A.2 Identification with two auxiliary measurements

Next, we expand our condensed model to include two continuous measurements. We show in this case, without the factor structure, $\alpha_0$ is not identified. This is in contrast with the identification result established in Theorem 3.1.

Suppose in addition to (A.1), we also observe two continuous measurements denoted as $Y_3$ and $Y_4$.

**Assumption 2.**
*1.* $(X_1, X) \perp (U, V, Y_3, Y_4)$.

*2.* $(X_1, X)$ *are continuously distributed with absolute continuous joint density w.r.t. Lebesgue measure. The conditional support of* $X_1$ *given* $X$ *is* $[a, b]$.
*3.* $V$ *is continuously distributed over* $\Re$ *and its density w.r.t. Lebesgue measure exist.*

**Theorem A.3.** *If Assumption 2 holds, then* $|\alpha_0| \leq b - a$ *is necessary and sufficient for* $\alpha_0$ *to be identified.*

The proof of Theorem A.3 is similar to that of Theorem A.1, and thus, is omitted. In the proof of Theorem A.1, we show that when $|\alpha_0| > b - a$, we can find an impostor $\alpha \neq \alpha_0$ and $\tilde{U}$ such that for any $x_1 \in [a, b]$ and any $v \in \text{Supp}(V)$, we have

$$P(\tilde{U} \leq x_1 + \alpha | V = v) = P(U \leq x_1 + \alpha_0 | V = v)$$
$$P(\tilde{U} \leq x_1 | V = v) = P(U \leq x_1 | V = v).$$

This implies the conditional CDF of $(Y_1, Y_2)$ given $(X_1, X)$ under the DGPs $(U, V, \alpha_0)$ and $(\tilde{U}, V, \alpha)$ are the same, and thus, $\alpha_0$ is observationally equivalent to the impostor $\alpha$. Similarly, with the two continuous measurements, we can use the exact same construction of $\tilde{U}$ and $\alpha$ to show that, for any $x_1 \in [a, b]$ and $(v, y_3, y_4) \in \text{Supp}(V, Y_3, Y_4)$, we have

$$P(\tilde{U} \leq x_1 + \alpha | V = v, Y_3 = y_3, Y_4 = y_4) = P(U \leq x_1 + \alpha_0 | V = v, Y_3 = y_3, Y_4 = y_4)$$
$$P(\tilde{U} \leq x_1 | V = v, Y_3 = y_3, Y_4 = y_4) = P(U \leq x_1 | V = v, Y_3 = y_3, Y_4 = y_4).$$

This implies the conditional CDF of $(Y_1, Y_2, Y_3, Y_4)$ given $(X_1, X)$ under the DGPs $(U, V, Y_3, Y_4, \alpha_0)$ and $(\tilde{U}, V, Y_3, Y_4, \alpha)$ are the same too. Such non-identification result holds even when $X$ has full support.

# B 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 A, to simplify exposition, in the following we focus exclusively on the parameters $\alpha_0, \gamma_0$. 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

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

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:

$$\hat{\theta} = \arg\max_{\theta} Q_n(\theta), \quad Q_n(\theta) \equiv \sum_{i \neq j} \hat{g}_{i,j}(\theta) \tag{B.2}$$

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** $\theta_0$ lies in the interior of $\Theta$, 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 (B.2) 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 F of the Supplementary Appendix which characterizes the rate of convergence and asymptotic distribution of the proposed estimator:

**Theorem B.1.** *Under Assumptions* **RK1-RK5**,

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, V^{-1}\Delta V^{-1}) \tag{B.3}$$

*where the forms of the Hessian term $V$ and outer score term $\Delta$ are described in detail in Section F of the Supplementary Appendix.*

# C 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

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

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

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, $(X_1, X) \perp (V, \Pi)$, $U = \gamma_0 V + \Pi$. $(V, \Pi)$ are distributed independently of each other, where $\Pi$ 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 $\sigma_x$ 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 $V$, $\alpha = 0.5$

| $\Pi$ | 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, 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 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 (C.1) and (C.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 $\sigma_{x1}$ and $\sigma_x$ 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.

# D Proof of Theorem 2.1

**Proof:** Note that

$$P^{11}(z_1, z_3, x) = \int_{-\infty}^{x} F_\Pi(z_1'\lambda_0 + z_3'\beta_0 + \alpha_0 - \gamma_0 v) f_V(v) dv$$
$$P^{10}(\tilde{z}_1, \tilde{z}_3, \tilde{x}) = \int_{\tilde{x}}^{+\infty} F_\Pi(\tilde{z}_1'\lambda_0 + \tilde{z}_3'\beta_0 - \gamma_0 v) f_V(v) dv.$$

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

$$\partial_x P^{11}(z_1, z_3, x)/f_V(x) = F_\Pi(z_1'\lambda_0 + z_3'\beta_0 + \alpha_0 - \gamma_0 x)$$
$$-\partial_x P^{10}(\tilde{z}_1, \tilde{z}_3, \tilde{x})/f_V(\tilde{x}) = F_\Pi(\tilde{z}_1'\lambda_0 + \tilde{z}_3'\beta_0 - \gamma_0 \tilde{x}).$$

By Assumption **A4**, we know that there exists pairs such that

$$Z_1'\lambda_0 + Z_3'\beta_0 + \alpha_0 - \gamma_0 X = \tilde{Z}_1'\lambda_0 + \tilde{Z}_3'\beta_0 - \gamma_0 \tilde{X}.$$

Because $F_\Pi(\cdot)$ is monotone increasing, we have

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

Note the LHS of the above display is identified from data. Denote $Z_{1,1}$ as the first element of $Z_1$, whose coefficient is set to one. The rest of $Z_1$ is denoted as $Z_{1,-1}$, whose coefficient is denoted as $\lambda_{0,-1}$. Then, we have

$$\alpha_0 + (Z_{1,-1} - \tilde{Z}_{1,-1})'\lambda_{0,-1} + (Z_3 - \tilde{Z}_3)'\beta_0 - \gamma_0(X - \tilde{X}) = \tilde{Z}_{1,1} - Z_{1,1}.$$

Then, by Assumption **A4**, we can find $(z_1^{(l)}, z_3^{(l)}, x^{(l)})_{l=1}^d$ and $(\tilde{z}_1^{(l)}, \tilde{z}_3^{(l)}, \tilde{x}^{(l)})_{l=1}^d$ such that

$$\operatorname{rank} \begin{pmatrix} 1 & \cdots & 1 \\ z_{1,-1}^{(1)} - \tilde{z}_{1,-1}^{(1)} & \cdots & z_{1,-1}^{(d)} - \tilde{z}_{1,-1}^{(d)} \\ z_3^{(1)} - \tilde{z}_3^{(1)} & \cdots & z_3^{(d)} - \tilde{z}_3^{(d)} \\ x^{(1)} - \tilde{x}^{(1)} & \cdots & x^{(d)} - \tilde{x}^{(d)} \end{pmatrix} = d.$$

Then, we can identify $(\alpha_0, \lambda_0, \beta_0, \gamma_0)$ by solving the linear system that

$$\begin{aligned} \alpha_0 + (z_{1,-1}^{(1)} - \tilde{z}_{1,-1}^{(1)})'\lambda_{0,-1} + (z_3^{(1)} - \tilde{z}_3^{(1)})'\beta_0 - \gamma_0(x^{(1)} - \tilde{x}^{(1)}) =& \tilde{z}_{1,1}^{(1)} - z_{1,1}^{(1)}, \\ &\vdots \\ \alpha_0 + (z_{1,-1}^{(d)} - \tilde{z}_{1,-1}^{(d)})'\lambda_{0,-1} + (z_3^{(d)} - \tilde{z}_3^{(d)})'\beta_0 - \gamma_0(x^{(d)} - \tilde{x}^{(d)}) =& \tilde{z}_{1,1}^{(d)} - z_{1,1}^{(d)}. \end{aligned}$$

This concludes the proof.

# E Proof of Theorem 3.1

For notation simplicity, we write $\tilde{W} = \nu_0 W$, $\tilde{\sigma}_0 = \sigma_0/\nu_0$, $\tilde{\nu}_0 = 1/\nu_0$, and

$$\begin{aligned} Y_2 &= \mathbf{1}\{X \geq \tilde{\nu}_0 \tilde{W} + \eta_2\} \\ Y_3 &= \tilde{W} + \eta_3 \\ Y_4 &= \tilde{\sigma}_0 \tilde{W} + \eta_4. \end{aligned}$$

Because Assumptions B2–B6 hold, by applying Hu and Schennach (2013, Theorem 1) to $Y_3$ and $Y_4$, we can identify the densities for $\nu_0 W = \tilde{W}$, $\eta_3$, and $\eta_4$ as well as $\sigma_0/\nu_0 = \tilde{\sigma}_0$.

Then, we have

$$\begin{aligned}\partial_{y_3}\mathbb{P}(Y_2 = 1, Y_3 \leq y_3|X = x) =& \partial_{y_3}\int F_{\eta_2}(x - \tilde{\nu}_0 w)F_{\eta_3}(y_3 - w)f_{\tilde{W}}(w)dw \\ =& \int F_{\eta_2}(x - \tilde{\nu}_0 w)f_{\eta_3}(y_3 - w)f_{\tilde{W}}(w)dw.\end{aligned}$$

Applying Fourier transform w.r.t. $y_3$ on both sides, we have

$$\mathcal{F}(\partial_{y_3}\mathbb{P}(Y_2 = 1, Y_3 \leq \cdot|X = x))(t) = \mathcal{F}(F_{\eta_2}(x - \tilde{\nu}_0\cdot)f_{\tilde{W}}(\cdot))(t)\mathcal{F}(f_{\eta_3}(\cdot))(t),$$

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

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

Therefore,

$$\frac{\mathcal{F}^{-1}\left(\frac{\mathcal{F}(\partial_{y_3}\mathbb{P}(Y_2=1,Y_3\leq\cdot|X=x))(\cdot)}{\mathcal{F}(f_{\eta_3}(\cdot))(\cdot)}\right)(w)}{f_{\tilde{W}}(w)} = F_{\eta_2}(x - \tilde{\nu}_0 w), \tag{E.1}$$

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

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

Note the LHS of (E.1) can be identified from data. We choose two pairs $(x, w)$ and $(x', w')$ such that $w \neq w'$ and

$$\frac{\mathcal{F}^{-1}\left(\frac{\mathcal{F}(\partial_{y_3}\mathbb{P}(Y_2=1,Y_3\leq\cdot|X=x))(\cdot)}{\mathcal{F}(f_{\eta_3}(\cdot))(\cdot)}\right)(w)}{f_{\tilde{W}}(w)} = \frac{\mathcal{F}^{-1}\left(\frac{\mathcal{F}(\partial_{y_3}\mathbb{P}(Y_2=1,Y_3\leq\cdot|X=x'))(\cdot)}{\mathcal{F}(f_{\eta_3}(\cdot))(\cdot)}\right)(w')}{f_{\tilde{W}}(w')}.$$

Then, given the monotonicity of $F_{\eta_2}$, we have

$$x - \tilde{\nu}_0 w = x' - \tilde{\nu}_0 w',$$

or

$$\tilde{\nu}_0 = (x - x')/(w - w'),$$

which is identified. Given the identification of $\tilde{\nu}_0$ and the distribution of $\tilde{W}$, we can identify the distribution of $W = \tilde{\nu}_0\tilde{W}$. Recall $F_{\eta_1}(\cdot)$ and $f_{\eta_2}(\cdot)$ are the CDF and PDF of $\eta_1$ and $\eta_2$, respectively.

Then, we have

$$P(Y_2 = 1|X = x) = P(W + \eta_2 \leq x).$$

Because $X$ has full support, we can identify the distribution of $W + \eta_2$. Then, it follows from standard deconvolution argument and the fact that the distribution of $W$ is identified that we can identify the distribution of $\eta_2$. In addition, note that

$$\begin{aligned} P^{11}(z_1, z_3, x) =& P(Y_1 = 1, Y_2 = 1|Z_1 = z_1, Z_3 = z_3, X = x) \\ =& \int F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 + \alpha_0 - \gamma_0 w) F_{\eta_2}(x - w) f_W(w) dw \end{aligned}$$

and

$$\begin{aligned} P^{10}(z_1, z_3, x) =& P(Y_1 = 1, Y_2 = 0|Z_1 = z_1, Z_3 = z_3, X = x) \\ =& \int F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 - \gamma_0 w)(1 - F_{\eta_2}(x - w)) f_W(w) dw. \end{aligned}$$

Taking derivatives of $P^{11}(z_1, z_3, x)$ and $P^{10}(z_1, z_3, x)$ w.r.t. $x$, we have

$$\partial_x P^{11}(z_1, z_3, x) = \int F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 + \alpha_0 - w) f_{\eta_2}(x - w) f_W(w) dw \tag{E.2}$$

and

$$-\partial_x P^{10}(z_1, z_3, x) = \int F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 - \gamma_0 w) f_{\eta_2}(x - w) f_W(w) dw. \tag{E.3}$$

Applying Fourier transform on both sides of (E.2) and (E.3), we have

$$\mathcal{F}(\partial_x P^{11}(z_1, z_3, \cdot)) = \mathcal{F}(F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 + \alpha_0 - \cdot) f_W(\cdot)) \mathcal{F}(f_{\eta_2}(\cdot)) \tag{E.4}$$

and

$$\mathcal{F}(-\partial_x P^{10}(z_1, z_3, \cdot)) = \mathcal{F}(F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 - \cdot) f_W(\cdot)) \mathcal{F}(f_{\eta_2}(\cdot)).$$

Then, by (E.4), we can identify $F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 + \alpha_0 - \cdot)$ by

$$F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 + \alpha_0 - \gamma_0 \cdot) = \mathcal{F}^{-1}\left( \frac{\mathcal{F}(\partial_x P^{11}(z_1, z_3, \cdot))}{\mathcal{F}(f_{\eta_2}(\cdot))} \right)(\cdot) / f_W(\cdot).$$

Similarly, we can identify

$$F_{\eta_1}(z_1'\lambda_0 + z_3'\beta_0 - \gamma_0 \cdot) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(-\partial_x P^{10}(z_1, z_3, \cdot))}{\mathcal{F}(f_{\eta_2}(\cdot))}\right)(\cdot)/f_W(\cdot).$$

Because $F_{\eta_1}(\cdot)$ is monotone increasing, we have

$$\begin{aligned}
&\mathcal{F}^{-1}\left(\frac{\mathcal{F}(\partial_x P^{11}(z_1, z_3, \cdot))}{\mathcal{F}(f_{\eta_2}(\cdot))}\right)(w)/f_W(w) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(-\partial_x P^{10}(\tilde{z}_1, \tilde{z}_3, \cdot))}{\mathcal{F}(f_{\eta_2}(\cdot))}\right)(\tilde{w})/f_W(\tilde{w}) \\
\iff &\alpha_0 + (z_1 - \tilde{z}_1)'\lambda_0 + (z_3 - \tilde{z}_3)'\beta_0 - \gamma_0(w - \tilde{w}) = 0
\end{aligned}$$

Then, by Assumption B7, we can find $(z_1^{(l)}, z_3^{(l)}, w^{(l)})_{l=1}^d$ and $(\tilde{z}_1^{(l)}, \tilde{z}_3^{(l)}, \tilde{w}^{(l)})_{l=1}^d$ such that

$$\operatorname{rank}\begin{pmatrix} 1 & \cdots & 1 \\ z_{1,-1}^{(1)} - \tilde{z}_{1,-1}^{(1)} & \cdots & z_{1,-1}^{(d)} - \tilde{z}_{1,-1}^{(d)} \\ z_3^{(1)} - \tilde{z}_3^{(1)} & \cdots & z_3^{(d)} - \tilde{z}_3^{(d)} \\ w^{(1)} - \tilde{w}^{(1)} & \cdots & w^{(d)} - \tilde{w}^{(d)} \end{pmatrix} = d.$$

Then, we can identify $(\alpha_0, \lambda_0, \beta_0, \gamma_0)$ by solving the linear system that

$$\begin{aligned}
\alpha_0 + (z_{1,-1}^{(1)} - \tilde{z}_{1,-1}^{(1)})'\lambda_{0,-1} + (z_3^{(1)} - \tilde{z}_3^{(1)})'\beta_0 - \gamma_0(w^{(1)} - \tilde{w}^{(1)}) &= \tilde{z}_{1,1}^{(1)} - z_{1,1}^{(1)}, \\
&\vdots \\
\alpha_0 + (z_{1,-1}^{(d)} - \tilde{z}_{1,-1}^{(d)})'\lambda_{0,-1} + (z_3^{(d)} - \tilde{z}_3^{(d)})'\beta_0 - \gamma_0(w^{(d)} - \tilde{w}^{(d)}) &= \tilde{z}_{1,1}^{(d)} - z_{1,1}^{(d)}.
\end{aligned}$$

This concludes the proof.

# F Proof of Theorem B.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

$$\begin{aligned}g_{i,j}(\theta) &= [\mathbf{1}\{\partial_2P^{11}(X_{1,i},X_i)/f_V(X_i)+\partial_2P^{10}(X_{1,j},X_j)/f_V(X_j)\geq 0\}\mathbf{1}\{\Phi(X_{1,i},X_i,X_{1,j},X_j;\theta)\geq 0\} \\ &+ \mathbf{1}\{\partial_2P^{11}(X_{1,i},X_i)/f_V(X_i)+\partial_2P^{10}(X_{1,j},X_j)/f_V(X_j)< 0\}\mathbf{1}\{\Phi(X_{1,i},X_i,X_{1,j},X_j;\theta)< 0\}],\end{aligned}$$

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

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

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 $\theta$ 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 $\theta_0$ 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 $\delta_n$ 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 F.1.** *(From Theorem 1 in Sherman (1994b)).*

*If $\delta_n$ and $\varepsilon_n$ are sequences of positive numbers converging to 0, and*

1. $\hat{\theta} - \theta_0 = o_p(\delta_n)$
2. *There exists a neighborhood of $\theta_0$ and a constant $\kappa > 0$ such that $\mathcal{G}(\theta) - \mathcal{G}(\theta_0) \geq \kappa\|\theta - \theta_0\|^2$*

*for all $\theta$ in this neighborhood.*

3. *Uniformly over $O_p(\delta_n)$ neighborhoods of $\theta_0$*

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

*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 F.2.** *(From Theorem 2 in Sherman (1994b)). Suppose $\hat{\theta}$ is $\sqrt{n}$-consistent for $\theta_0$, an interior point of $\Theta$. Suppose also that uniformly over $O_p(n^{-1/2})$ neighborhoods of $\theta_0$,*

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

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

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

We first turn attention to applying Theorem F.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 F.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 $\theta$, $\mathcal{G}(\theta)$ 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

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

where $\nabla_\theta$ and $\nabla_{\theta\theta}$ denote first and second derivative operators and $\theta^*$ 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

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

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

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

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

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

which suffices for the second condition to hold.

To show the third condition in Theorem F.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

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

and

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

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

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

and

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

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:

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

An immediate technical difficulty that arises with the above term is the presence of a nonpara-

metric 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 (F.15) can be represented as:

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

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 $\theta$ lying in $o_p(1)$ neighborhoods of $\theta_0$. Therefore, uniformly for $\theta$ in an $o_p(1)$ neighborhood of $\theta_0$, 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 F.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:

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

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 F.17 can be expressed as:

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

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

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:

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

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 $\theta$. 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 (F.19), which is of the form:

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

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

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

Note that by Assumptions **RK2**, **RK3**, $\mathcal{H}(x_i,\theta)$ is smooth in $\theta$ 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 (F.22) as the which we note is a mean 0 sum

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

where

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

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

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

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

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

where $\psi_{2rnki}$ 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 (F.16) can be represented as:

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

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

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

where

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

and

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

# G Model with Two Idiosyncratic Shocks

In this section, we focus on the identification of $(\alpha_0,\gamma_0)$ in the "condensed" model that $X_1=Z_1'\lambda_0+Z_3'\beta_0$ is observed and

$$\begin{aligned} Y_1&=\mathbf{1}\{X_1+\alpha_0 Y_2-U\geq 0\}\\ Y_2&=\mathbf{1}\{X-V\geq 0\}. \end{aligned} \tag{G.31}$$

with the understanding that $(\lambda_0,\beta_0)$ can be identified jointly with $\alpha_0$ and $\gamma_0$, as shown in Theorems 2.1 and 3.1. We further impose $U=\gamma_0 W+\eta_1$, $V=W+\eta_2$, and $(W,\eta_1,\eta_2)$ 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 $\alpha_0$. But identification of $\alpha_0$ becomes more difficult in this model without the help of repeated measurements, as established in the following theorem.

**Theorem G.1.** *Suppose* (G.31) *holds,* $\gamma_0$ *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* $\alpha_0$ *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** $W$, $\eta_1$ and $\eta_2$ are mutually independent.

**C4** $X$ has full support.

**C5** Denote the density of $\eta_2$ as $f_{\eta_2}$, then $f_{\eta_2}$ 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 $\phi_\sigma$ is the density for a normal distribution with zero mean and $\sigma^2$ variance.

Assumption **C5** effectively assumes that the distribution of $\eta_2$ 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 $\sigma_0$ and the density of $\eta_2$. If $\eta_2$ 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 $\eta_2'$ and $W$ and $\tilde{\sigma} > 0$. It implies

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

where $\eta_2'$ 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**, $\sigma_0$ is not identified.

**Theorem G.2.** *If Assumptions* ***C1–C5*** *hold, then* $\sigma_0$*,* $\gamma_0$ *and* $\alpha_0$ *are 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 $\sigma_0$ is known, then Assumption **C5** is not necessary.

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.

# H Proof of Theorem A.1

Denote $P^{ij}(x_1, x) = Prob(Y_1 = i, Y_2 = j | X_1 = x_1, X = x)$. Then

$$\begin{aligned} P^{11}(x_1, x) &= \int_{-\infty}^{x} F_U(x_1 + \alpha_0 | V = v) f(v) dv \\ P^{10}(\tilde{x}_1, x) &= \int_{x}^{+\infty} F_U(\tilde{x}_1 | V = v) f(v) dv. \end{aligned} \tag{H.32}$$

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

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

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

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

This concludes the sufficient part.

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

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

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

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

where the third equality holds because, since $\alpha_0 < a - b$ and $\alpha < \alpha_0$, $b + \alpha_0 \leq x_1 + \alpha_0 - \alpha$ for $x_1 \in [a, b]$. Let $G_{U,V}$ and $G_{\tilde{U},V}$ be the joint distribution of $(U, V)$ and $(\tilde{U}, V)$ respectively. Then the above calculation with (H.32) imply that $(\alpha_0, G_{U,V})$ and $(\alpha, G_{\tilde{U},V})$ produce the identical pair $(P^{11}(x_1, x), P^{10}(x_1, x))$. In addition, the distribution of $V$ is unchanged so that $P(Y_2 = 1 | X = x)$ is identified from data. Therefore, $(\alpha_0, G_{U,V})$ and $(\alpha, G_{\tilde{U},V})$ are observationally equivalent.

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

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

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

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

where we use the facts that $x_1 \leq a + \alpha_0$ and $x_1 - a < \alpha$ for $x_1 \in [a, b]$. So again, $(\alpha_0, G_{U,V})$ and $(\alpha, G_{\tilde{U},V})$ are observationally equivalent.

# I Proof of Theorem A.2

The sign of $\alpha_0$ is identified by the data. In the following, we focus on deriving the results when $\alpha_0 > b - a$. By the proof of Theorem A.1, we have already shown that all $\alpha > \alpha_0$ is in the identified

set. Now we consider $\frac{b-a+\alpha_0}{2} \leq \alpha < \alpha_0$.

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

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

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

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

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

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

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

Now repeating the above construction but replacing $U$ with $U^{(1)}$ and $\alpha_0$ with $\alpha^{(1)}$, we have for any $\alpha^{(1)} > \alpha \geq \alpha^{(2)} \equiv \frac{b-a+\alpha^{(1)}}{2}$, there exists $U^{(2)}(\alpha)$ such that for any $x_1 \in [a, b]$, any $v$ and any $\alpha^{(1)} > \alpha \geq \alpha^{(2)}$,

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

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

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

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

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

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

# J Proof of Theorem G.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 $\alpha_0$ is not identified if $\alpha_0 > b-a$. To establish this, consider an impostor $\alpha$ 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 $\alpha$ exists because of the fact that $1-(b-a) > \alpha_0 > b-a$. Let $\Delta = \alpha_0 - \alpha$ and $(\tilde{W}, \tilde{\eta}_1, \tilde{\eta}_2)$ be mutually independent such that $\tilde{W}$ is distributed as $W-\Delta$, $\tilde{\eta}_2$ is distributed as $\eta_2 - \Delta$, and

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

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

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

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

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

This implies $\alpha_0$ is not identified from the impostor $\alpha$.

# K Proof of Theorem G.2

We first show that both $\sigma_0$ and the density of $\eta_2$ 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 $\sigma_0$ are not identified so that there exist $f'_{\eta_2}(\cdot)$ and $\sigma'$ 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 $\sigma_0$ are known. Recall $F_{\eta_1}(\cdot)$ as the CDF of $\eta_1$. 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

$$\partial_x P^{11}(x_1,x) = \int F_{\eta_1}(x_1+\alpha_0-\gamma_0 w) f_{\eta_2}(x-\sigma_0 w) f_W(w)dw \tag{K.33}$$

and

$$-\partial_x P^{10}(x_1,x) = \int F_{\eta_1}(x_1-\gamma_0 w) f_{\eta_2}(x-\sigma_0 w) f_W(w)dw. \tag{K.34}$$

Applying Fourier transform on both sides of (K.33) and (K.34), we have

$$\mathcal{F}(\partial_x P^{11}(x_1,\cdot)) = \mathcal{F}_{\sigma_0}(F_{\eta_1}(x_1+\alpha_0-\gamma_0\cdot)f_W(\cdot))\mathcal{F}(f_{\eta_2}(\cdot)) \tag{K.35}$$

and

$$\mathcal{F}(-\partial_x P^{10}(x_1,\cdot)) = \mathcal{F}_{\sigma_0}(F_{\eta_1}(x_1-\gamma_0\cdot)f_W(\cdot))\mathcal{F}(f_{\eta_2}(\cdot)), \tag{K.36}$$

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 (K.35), 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 $\alpha_0$ and $\gamma_0$ as the solution of the following linear system:

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

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

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).

Hu, Y., and S. M. Schennach (2013): "Nonparametric identification and semiparametric estimation of classical measurement error models without side information," *Journal of the American Statistical Association*, 108(501), 177–186.

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

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

Reiersol, O. (1950): "Identifiability of a Linear Relation Between Variables Which are Subject to Error," *Econometrica*, 18(4), 375–389.

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

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

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

Vytlacil, E. J., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75(3), 757–779.