Inference on Semiparametric Multinomial Response Models*

Shakeeb Khan[†] Boston College

Fu Ouyang[‡] University of Queensland

Elie Tamer[§] Harvard University

May 12, 2019

Abstract

In this paper we explore inference on regression coefficients in semi parametric multinomial response models. We consider cross sectional, and both static and dynamic panel settings where we focus throughout on point inference under sufficient conditions for point identification. The approach to identification uses a matching insight throughout all three models and relies on variation in regressors: with cross section data, we match across individuals while with panel data we match within individuals over time. Across models, IIA is not assumed as the unobserved errors across choices are allowed to be arbitrarily correlated. For the cross sectional model estimation is based on a localized rank objective function, analogous to that used in Abrevaya, Hausman, and Khan (2010), and presents a generalization of existing approaches. In panel data settings rates of convergence are shown to exhibit a curse of dimensionality in the number of alternatives. The results for the dynamic panel data model generalizes the work of Honoré and Kyriazidou (2000) to cover the multinomial case. A simulation study establishes adequate finite sample properties of our new procedures and we apply our estimators to a scanner panel data set.

Keywords: Multinomial choice, Rank Estimation, Adaptive Inference, Dynamic Panel Data.

JEL: C22,C23,C25.

*We are grateful for helpful comments received from conference participants at the 2016 Panel Data Work Shop at University of Amsterdam, the 22nd Panel Data Conference in Perth, Australia, the 2016 Australasian Meetings of the Econometric Society at UTS, and seminar participants at various institutions.

[†]Email: shakeeb.khan@bc.edu.

[‡]Email: f.ouyang@uq.edu.au.

[§]Email: elietamer@fas.harvard.edu.

1 Introduction

Many important economic decisions involve households' or firms' choice among qualitative or discrete alternatives. Examples are individuals' choice among transportation alternatives, family sizes, residential locations, brands of automobile, health plans etc. The theory of discrete choice is designed to model these kinds of choice settings and to provide the corresponding econometric methodology for empirical analyses. Due to variables that are unobservable to the econometrician, the observations from a sample of agents' discrete choices can be viewed as outcomes generated by a stochastic utility maximization model. In the context of choice behavior, the probabilities in the multinomial model are to be interpreted as the probability of choosing the respective alternatives (choice probabilities) and so one is interested in expressing the choice probabilities as functions of the agents' preferences and the choice constraints. As in most of the econometrics literature, agents know their own utilities and make a choice from a well defined choice set, while the econometrician knows the choice set, observes choices and covariates and the objective here is to learn the finite dimensional coefficients that would characterize utilities of various alternatives.

There has been a renewed interest recently among applied economists in estimating models of multinomial choice with both cross section and panel data. In the marketing, IO and other literatures, recent papers have also emphasized the role of dynamics in panel data settings. See for example Merlo and Wolpin (2015) for an application to a dynamic model of schooling, and crime, Handel (2013) for a model of health insurance choice among others.[1] A central question in these models is the separation of heterogeneity from state dependence. More broadly in econometric theory, there has been a push for semiparametric work in models that relax the IIA assumption in both cross section and panel data setups. For example, Ahn, Powell, Ichimura, and Ruud (2017) studies this problem with cross section data, Pakes and Porter (2014) and Shi, Shum, and Song (2018) study multinomial panel data models without IIA, while Khan, Ponomareva, and Tamer (2019) analyze the identification question in binary response models in dynamic panels under weak assumptions. More recently Gao and Li (2019) provide novel identification results in panel multinomial models when the link function can be unknown and/or nonseparable in the fixed effects.

In this paper, we focus on inference on cross sessional and panel data multinomial response models under point identification where we use a unified approach for identification in all three classes of multinomial models: cross sectional, static panels, and dynamic

[1]Other interesting papers include Dubé, Hitsch, and Rossi (2010), Illanes (2016), Ketcham, Lucarelli, and Powers (2015), Polyakova (2016), Raval and Rosenbaum (2018).

panels. Throughout we relax the IIA property by allowing for arbitrary correlation in the unobserved errors across choices. In cross sectional settings, we match different individuals or units in a particular way to obtain a monotone index model that is familiar in econometrics. This matching requirement is guaranteed to hold under the conditions we require on the regressors. We then generalize this matching approach to static panel data settings and require different variation in the regressors over time to garner point identification. Again, here, the contribution is a model that generalizes Manski (1987) to allow for correlation in the unobservables (and hence relax the IIA property that he maintains). We derive rates of convergence in the multinomial maximum score estimator and show that it is a function of the *number of alternatives*. Finally, we provide point identification results in the panel multinomial model with dynamics, provide an estimator in this case and study its rate of convergence. This generalizes the work of Honoré and Kyriazidou (2000) to multinomial settings and complements their work by providing large sample rates of convergence for the various estimators. Our approach is robust in the sense that it achieves meaningful bounds for the regression coefficients when conditions for point identification fail, such as when all the regressors are discrete.

We structure the paper as follows. In the next section we formally introduce the cross-sectional model, and state standard regularity conditions on both observed and unobserved random variables that guarantee point identification. This model introduces the main intuition for how we get identification in this paper and can be clearly explained. This identification strategy also motivates a localized rank based objective function. We then show that this model yields a root-n consistent and asymptotically normal estimator under appropriate conditions.

Section 3 generalizes the cross sectional model by assuming the availability of a longitudinal panel data set and introducing unobserved individual and choice specific effects. For this model we propose a localized maximum score (similar to Manski (1987)) estimator and show that and under certain DGPs is point consistent. Most interestingly in this paper, we further generalize the multinomial model by introducing dynamics in Section 3.2. Specifically, we do so by allowing lagged values of dependent variables to be explanatory variables. This approach of modeling dynamics was taken in the binary choice model. See, e.g., Heckman (1978), Honoré and Kyriazidou (2000), Chen, Khan, and Tang (2015), and Khan et al. (2019). Here again, our procedure is shown to be point consistent under standard conditions.

Section 4 explores finite sample properties of the new procedures through a small scale simulation study and Section 5 applies the new procedures using an optical scanner panel data set on purchase decisions in the saltine cracker market. Section 6 concludes by summarizing results and proposing areas for future research. The appendix collects

proofs of many of the theorems stated in the paper.

2 Semiparametric Multinomial Choice

We consider the standard multinomial choice model where the dependent variable takes one of $J+1$ mutually exclusive and exhaustive alternatives (numbered from $j=0$ to $j=J$). Specifically, for individual i, alternative j is assumed to have an unobservable indirect utility y_{ij}^* for that individual. The alternative with the highest indirect utility is assumed chosen. Thus the observed variable y_{ij} has the form

$$y_{ij} = \mathbf{1}[y_{ij}^* > y_{ik}^* \text{ for all } k \neq j]$$

with the convention that $y_{ij}=0$ indicates choice of alternative j is not made by agent i. As is standard in this literature an assumption of joint continuity of the indirect utilities rules out ties (with probability one). In addition, we maintain the assumption that the indirect utilities are restricted to have the linear form

$$\begin{aligned} y_{i0}^* &= 0 \\ y_{ij}^* &= x_{ij}'\beta_0 - \epsilon_{ij},\ j=1,...,J \end{aligned}$$

where β_0 is a (p)-dimensional vector of unknown parameters of interest whose first component is normalized to have absolute value 1 (scale normalization). Note that for alternative $j=0$, the standard (location) normalization $y_{i0}^*=0$ is imposed. The vector $\epsilon_i=(\epsilon_{i1},...,\epsilon_{iJ})$ of unobserved error terms attained by stacking all the scalars ϵ_{ij}, is assumed to be jointly continuously distributed and independent of the $J\times p$- dimensional matrix of regressors X_i whose j^{th} row is x_{ij}'.

Parametric assumptions on the unobservables ϵ_i can be maintained such as an iid Type 1 extreme value (multinomial Logit) or multivariate normal (multinomial Probit). The multinomial Logit model suffers from the well known IIA problem (McFadden (1978)). The multinomial Probit on the other hand leads to choice probabilities that are difficult to compute. There has been approaches to ameliorate both problems by for example using nested logit models, and simulation approaches have been successfully used to approximate multiple integrals.

We take another approach. We are interested in the question of what is required to point identify β_0 when minimal assumptions are made on the joint distribution of ϵ_i. The identification here is first motivated by the discussion in Khan and Tamer (2018) where the matching was briefly explored. Previous contributions to this question include Lee

(1995) who imposed a profile likelihood approach, extending the results in Klein and Spady (1993) for the binary choice model. Ahn et al. (2017) propose a 2-step estimator that requires nonparametric methods but show the second stage is of closed form. Shi et al. (2018) also propose a 2-step estimator in panel setups exploiting a cyclic monotonicity condition, which also requires high dimensional nonparametric first stage, but whose second stage is not closed form as Ahn et al. (2017) is.

The next section contains the main intuition that runs through the various models in the paper. It is provided for the static multinomial model.

2.1 Local Rank Procedure

Consider a multinomial choice model with 3 choices ($J = 2$) for now where the latent utilities for alternatives 0, 1, and 2 are:

$$\begin{aligned} y_{i0}^* &= 0 \\ y_{i1}^* &= x_{i1}'\beta_0 - \epsilon_{i1} \\ y_{i2}^* &= x_{i2}'\beta_0 - \epsilon_{i2} \end{aligned}$$

with the maintained assumption that

$$(\epsilon_{i1}, \epsilon_{i2}) \perp (x_{i1}, x_{i2})$$

but we allow arbitrary correlation between ϵ_{i1} and ϵ_{i2}.

From a random sample of $(y_{i0}, y_{i1}, y_{i2}, x_{i1}, x_{i2})$ where

$$y_{ij} = \mathbf{1}[y_{ij}^* > y_{ik}^*, \forall k \neq j], \ j = 0, 1, 2$$

we are interested in bounds for β_0, even when all regressors have discrete support.

From the model and assumptions we have for a given β_0 and joint distribution F_ϵ on the ϵ_i, the choice probabilities are given by:

$$G(x_{i1}, x_{i2}; \beta_0, F_\epsilon) = \begin{bmatrix} P(x_{i1}'\beta_0 - \epsilon_{i1} \leq 0; x_{i2}'\beta_0 - \epsilon_{i2} \leq 0) \\ P(x_{i1}'\beta_0 - \epsilon_{i1} \geq 0; x_{i1}'\beta_0 - \epsilon_{i1} \geq x_{i2}'\beta_0 - \epsilon_{i2}) \\ P(x_{i1}'\beta_0 - \epsilon_{i1} \leq x_{i2}'\beta_0 - \epsilon_{i2}; x_{i2}'\beta_0 - \epsilon_{i2} \geq 0) \end{bmatrix}$$

To illustrate how we garner information about β_0 we first fix x_{i2} and illustrate with the choice probability for the first alternative. With x_{i2} fixed, we have what we call a *conditional monotone index model*. By this we mean that conditional on X_i, $P(y_{i1} = 1 | x_{i1}, x_{i2} = x_2)$

is increasing in $x_{i1}'\beta$ for all constant vector x_2. Thus for all $i \neq m$, we have the following relationship

$$P(y_{i1} = 1|x_{i1}, x_{i2} = x_2) \geq P(y_{m1} = 1|x_{m1}, x_{m2} = x_2) \iff x_{i1}'\beta_0 \geq x_{m1}'\beta_0$$

Fixing regressors of one alternatives to obtain a monotonic index models motivates all our identification results.

Note the above conditional moment inequalities can be repeated for all values of x_2 (finitely many if the support of x_{i2} is finite). Furthermore, note for a fixed x_2, $P(y_{i0} = 1|x_{i1}, x_{i2} = x_2)$ and $P(y_{i2} = 1|x_{i1}, x_{i2} = x_2)$ are both decreasing in $x_{i1}'\beta_0$. This can be exploited also fixing x_{i1} resulting in other moment inequalities. Collectively, all these moment inequalities can be used to study the conditions needed for identification.

This local monotonicity translates into an estimation procedure, which will converge to an informative region.[2] For example, assuming a random sample of n observations, we propose the following weighted[3] rank correlation estimator, analogous to the maximum rank correlation (MRC) estimator proposed in Han (1987), defined here as the maximizer over the parameter space $\mathcal{B}$, of the objective function

$$G_{1n}^{(1)}(b) = \frac{1}{n(n-1)} \sum_{i \neq m} \mathbf{1}[x_{i2} = x_{m2}] sgn(y_{i1} - y_{m1}) sgn((x_{i1} - x_{m1})'b) \tag{2.1}$$

where $sgn(\cdot)$ above denotes the sign operator. The above function can be used for one set of moment inequalities. But as alluded to, we can also work with y_{i0} and y_{i2}. In addition to these, we can fix x_{i1}, which can yield objective functions of the form

$$G_{2n}^{(2)}(b) = \frac{1}{n(n-1)} \sum_{i \neq m} \mathbf{1}[x_{i1} = x_{m1}] sgn(y_{i2} - y_{m2}) sgn((x_{i2} - x_{m2})'b)$$

It is clear any one or a combination of the above objective functions above can be used for inference on β.

Remark 1. *In the case where the regressors for choices 1 and 2 are the same ($x_1 \equiv x_2$), then the choice probability for choice 1 foer example is monotonic in the index and a standard rank estimator applies. In the more interesting case when there are some regressors in common such as $x_1\beta \equiv \tilde{x}_1\beta + z\gamma$ and $x_2\beta \equiv \tilde{x}_2\beta + z\gamma$, a two step procedure is possible to get inference on both β and γ. In the first step, we use the objective function above to get β and hence the indices*

[2]Note that when we are conditioning on, say x_{i2} being fixed yet allowing x_{i1} to vary we are implicitly assuming exclusion between components of these vectors. For more on strategies for inference when regressors are common across alternative, see Remark 1 below.

[3]Here the weights correspond to binary, "exact" matches of each component of the vector x_2.

$x_1^* = \tilde{x}_1\beta$ and $x_2^* = \tilde{x}_2\beta$. This objective function does not get us information about γ since with the matching, $z\gamma$ drops out. But, once x_1^* and x_2^* are "known", then one can use another rank procedure in a second step where we condition on $x_1^* = x_2^*$. The choice probability for choice 1 for example becomes:

$$P(1|x_1, x_2, z, x_1^* = x_2^*) = P(x_1^* + z\gamma + \epsilon_1 \geq 0; 0 \geq \epsilon_1 - \epsilon_2)$$

This is another version of the conditional monotone index model (monotone in $z\gamma$).

A key condition for point identification is the usual full support conditions for one of the regressors. This condition is stated next.

AS1: At least one of the vectors $x_1, x_2, ..., x_J$ has one component which is continuously distributed with positive density on the real line.

Such a support condition is analogous to that assumed in Manski (1975) and Han (1987).

Note that it is hard to match regressors if these are continuous,[4] thus the value of the objective function will always be 0. But here we can construct kernel weights as follows. To illustrate for the $G_{1n}^{(1)}(b)$ objective function, assuming the regressors for choice 2 have at least one continuous component, we construct the approximate binary weights with:

$$\mathbf{1}[x_{i2} = x_{m2}] \approx K_h(x_{i2} - x_{m2}) \equiv w_{im}$$

with $K_h(\cdot) = K(\cdot/h_n)$ where K is a kernel density function and h_n is a bandwidth sequence that converges to 0 as $n \to \infty$.

The following theorem establishes point identification for β_0 and limiting distribution theory for $\hat{\beta}$, the estimator defined as the maximizer of $G_{1n}^{(1)}(b)$ in 2.1. The proof is omitted as the same arguments used in Han (1987) and Sherman (1993) can be applied to this objective function.

Theorem 2.1. *In the multinomial choice model, assume $\epsilon_i \equiv (\epsilon_{i1}, ...\epsilon_{iJ})$ is distributed independently of $x_i = (x_{i1}, ...x_{iJ})$ with support on R^J. Furthermore, assume that for some $j \neq 1$, the first component of x_{i1} is continuously distributed with support on the real line, conditional on x_{ij}. Also, assume that conditional on x_{ij}, the vector x_{i1} does not lie in a linear subspace of R^p. Then β_0 is point identified.*

[4] But note this depends on the choice in question. For example, consider the same 3 choice setting. Suppose for choice 1, the regressor has one continuous component with support on the real line, but the other components for choice 1 regressors are discrete. Suppose for choice 2 all the components of the regressors are discrete. Then we can match as in $G_{1n}^{(1)}(b)$ and $G_{2n}^{(2)}(b)$, and this in fact will point identify β_0.

In addition, assume a random sample of observations of the vector $(y_i, x_i), i = 1, 2, ...n$, $G_{1n}^{(1)}(b)$ *converges uniformly in probability to* $G_1(b) \equiv E[G_{1n}^{(1)}(b)]$ *and* $\mathcal{B}$ *is a compact subset of* R^p. *Define our estimator as*

$$\hat{\beta} \equiv \arg\max_{b \in \mathcal{B}} G_{1n}^{(1)}(b) \tag{2.2}$$

Then $\hat{\beta} \xrightarrow{p} \beta_0$.

Furthermore, if $G_1(b)$ *is twice continuously differentiable in* b *for all* b *in a neighborhood of* β_0, *and its second derivative, a* $p \times p$ *matrix, is invertible at* $b = \beta_0$, *we have limiting distribution theory for the estimator defined in 2.2; to characterize it, let* ζ_i *denote the vector* y_{i1}, x_{i1}, x_{i2} *and define*

$$\tau(\zeta_i, b) = E_m \left[I[y_{i1} > y_{m1}] I[x_{i1}' b > x_{m1}' b] I[x_{i2} = x_{m2}] + I[y_{m1} > y_{i1}] I[x_{m1}' b > x_{i1}' b] I[x_{m2} = x_{i2}] \right]$$

where the operator $E_m[\cdot]$ *denotes the expectation taken with respect to* ζ_m; *then*

$$\sqrt{n}(\hat{\beta} - \beta_0) \Rightarrow N(0, V^{-1} \Sigma V^{-1}) \tag{2.3}$$

where

$$V \equiv \frac{1}{2} E[\nabla_2 \tau(\zeta_i, \beta_0)]$$

and

$$\Sigma \equiv E[\nabla_1 \tau(\zeta_i, \beta_0) \nabla_1 \tau(\zeta_i, \beta_0)'$$

where ∇_1, ∇_2 *denote first and second order derivative operators, respectively.*

For the case where the regressors used for matching are continuously distributed, so kernel weighting is required, we also have limiting distribution theory. For ease of illustration, again here we consider consider the case where $J = 3$. The objective function of $\beta \in \mathcal{B}$, is now of the form:

$$\frac{1}{n(n-1)} \sum_{i \neq m} K_h(x_{i2} - x_{m2}) I[y_i^{(1)} > y_m^{(1)}] I[x_{i1}' \beta > x_{m1}' \beta] \tag{2.4}$$

To characterize the limiting distribution theory for the estimator defined as the maximizer of 2.4, we will use the following notation: ζ_i denotes the vector $y_i^{(1)}, x_{i1}, x_{i2}$; define

$$\begin{aligned} \mu(y_i^{(1)}, x_{i1}, \beta, x_{j2}) &= E\Big[I[y_i^{(1)} > y_m^{(1)}] I[x_{i1}' \beta > x_{m1}' \beta] \\ &+ I[y_m^{(1)} > y_i^{(1)}] I[x_{m1}' \beta > x_{i1}' \beta] \Big| y_i^{(1)}, x_{i1}, x_{j2} \Big] \end{aligned} \tag{2.5}$$

Then define

$$\tau_2(\zeta_i, \beta) = \mu(y_i^{(1)}, x_{i1}, \theta, x_{i2}) \cdot f_2(x_{i2})$$

where $f_2(\cdot)$ denotes the density function of x_{i2}. The function $\tau_2(\cdot, \cdot)$ will characterize the limiting distribution of the maximizer of 2.4. We have the limiting distribution theorem theorem, whose proof follows from identical arguments to those used in Abrevaya et al. (2010), and is based on the following regularity conditions:

KWR1 The parameter space $\mathcal{B}$ is compact.

KWR2 Random sampling of $(y_i^{(1)}, x_{i1}, y_i^{(2)}, x_{i2})$

KWR3 $\epsilon_{i1}, \epsilon_{i2}$ is distributed independently of x_{i1}, x_{i2}, and has density function which is positive on R^2.

KWR4 Conditional on x_{i2}, x_{i1} has rank p.

KWR5 For all β in a neighborhood of β_0 and all ζ_i in its support, $\tau(\zeta_i, \beta)$ is twice continuously differentiable with respect to β.

KWR6 The $p \times p$ matrix $\nabla_2 \tau_2(\zeta_i, \beta_0)$ is invertible, where $\nabla_2 \tau_2(\cdot, \cdot)$ denotes the second derivative of $\tau_2(\cdot, \cdot)$ with respect to its second argument.

KWR7 The $p \times 1$ vector $\nabla_1 \tau_2(\zeta_i, \beta_0)$ has finite second moment, where $\nabla_1 \tau_2(\cdot, \cdot)$ denotes the first derivative of $\tau_2(\cdot, \cdot)$ with respect to its second argument.

KWR8 The density function of x_{2i}, $f_2(\cdot)$ is ℓ times continuously differentiable with bounded ℓ^{th} derivative, where ℓ is an even integer satisfying $\ell > p/2$.

KWR9 The kernel function $K(\cdot)$is of order ℓ and h_n satisfies $\sqrt{n} h_n^{\ell} \to 0$ and $n h_n^{p} \to \infty$.

Theorem 2.2. *Under Assumptions KWR1-KWR9,*

$$\sqrt{n}(\hat{\beta} - \beta_0) \Rightarrow N(0, V^{-1} \Delta V^{-1})$$

where $V = \frac{1}{2} E[\nabla_2 \tau_2(\zeta_i, \beta_0)$ *and* $\Delta = E[\nabla_1 \tau_2(\zeta_i, \beta_0) \nabla_1 \tau_2(\zeta_i, \beta_0)']$.

3 Panel Data Multinomial Choice

3.1 Static Multinomial Choice

Paralleling the increase in popularity of estimating multinomial response models in applied work is the estimation of panel data models. The increased availability of longitudinal panel data sets has presented new opportunities for econometricians to control for individual unobserved heterogeneity across agents. In linear panel data models, unobserved additive individual-specific heterogeneity, if assumed constant over time (i.e., "fixed effects"), can be controlled for when estimating the slope parameters by first differencing the observations.

Discrete panel data models have received a great deal of interest in both the econometrics and statistics literature, beginning with the seminal paper of Andersen (1970). For a review of the early work on this model see Chamberlain (1984), and for a survey of more recent contributions see Arellano and Honoré (2001). See also the key contribution in Manski (1987) on maximum score binary response model. More generally, there is a vibrant and growing literature on both partial and point identification in nonlinear panel data models. There are a set of recent papers that deal with various nonlinearities in models with short panels ($T < \infty$). See for example the work of Arellano and Bonhomme (2009), Bonhomme (2012), Chernozhukov, Val, Hahn, and Newey (2013), Graham and Powell (2012), Hoderlein and White (2012), Khan, Ponomareva, and Tamer (2016), and Chen et al. (2015). See also the survey in Arellano and Honoré (2001).

Here we consider a panel data model for multinomomial choice like Chamberlain (1984) where the latent utility and observed choices can be expressed as

$$y^*_{ijt} = x'_{ijt}\beta_0 + \alpha_{ij} - \epsilon_{ijt}$$

and

$$y_{it} = \arg \max_{0 \leq j \leq J} y^*_{ijt}$$

for $i = 1, ..., n$, $j = 0, 1, ..., J$, and $t = 1, ..., T$. Thus in our notation, for the subscript ijt the first component i denotes the individual, the second component j denotes the choice/product, the third component t denotes the time period. Note also that the utilities above include a set of *fixed effects* α_{ij} that are *both* individual and choice/product specific. Throughout, no assumptions are made on the distribution of $\alpha's$ conditional on x and ϵ.

In what follows, we denote $y_{ijt} = \mathbf{1}[y_{it} = j]$. We will consider identification and asymptotics with J, T fixed and $n \to \infty$. Existing results for panel data binary choice models with fixed effects include Andersen (1970), Manski (1987), and Chamberlain (2010)

among others. The literature on multinomial choice for panel data is more limited. Recent results include Shi et al. (2018) and Pakes and Porter (2014). The latter is concerned with partial identification. The former achieves point identification. For recent work on partial identification in binary dynamic panel data models under weak assumptions, see also Khan et al. (2019).

Here we propose point identification results under similar weak conditions as ones used in Manski (1987). To illustrate our identification results, assume $T = 2, J = 2$ (So the choice set is $\{0, 1, 2\}$, with 2 time periods) w.l.o.g. and as before impose normalization that $y^*_{i0t} \equiv 0$ for $t = 1, 2$.

Our identification strategy will be analogous to the cross-sectional case, but now we match and do our comparisons *within* individuals over time as opposed to pairs of individuals. As we will show the analogy is not perfect as we have to condition on "switchers", in a way similar to the estimation of the conditional Logit model in Andersen (1970) and the conditional maximum score estimator in Manski (1987). Besides that here we also need a subset of the population whose regressor values for a choice changes over time, but whose regressors for a *different* choice are time-invariant.

Specifically, one objective function we work with is[5]:

$$G_n^{SP}(b) = \sum_{i=1}^{n} \mathbf{1}[x_{i21} = x_{i22}](y_{i11} - y_{i12}) sgn\left((x_{i11} - x_{i12})'b\right) \tag{3.1}$$

Note that this objective function is turned off for observations where $y_{i11} = y_{i12}$, i.e., when individual i chooses alternative 1 in both periods 1 and 2. The objective function then uses only *switchers*, or individuals whose choice changes over time.

For the case when x_{i21}, x_{i22} has continuous components, we replace the indicator function with a kernel function,

$$G_n^{SP}(b) = \sum_{i=1}^{n} K_{h_n}(x_{i21} - x_{i22})(y_{i11} - y_{i12}) sgn\left((x_{i11} - x_{i12})'b\right) \tag{3.2}$$

where $K(\cdot)$ denotes a kernel function, and h_n denotes a bandwidth sequence. Under conditions analogous to Manski (1987), which we state below, β_0 is point identified and the maximizer of $G_n^{SP}(b)$ is a consistent estimator.[6] To facilitate exposition in stating our conditions, we first introduce the following notation.

[5] In addition to this objective function, one can use the objective function that use choices 0 and 2.

[6] As was the case in the cross sectional model, point identification is not attainable when all the regressors are discrete.

Notation: (i) $y_t \equiv (y_{0t}, y_{1t}, y_{2t})'$, $\epsilon_t \equiv (\epsilon_{1t}, \epsilon_{2t})'$, and $x_t \equiv (x_{1t}', x_{2t}')'$ for all t, where $x_{jt} \equiv (x_{jt}^{(1)}, ..., x_{jt}^{(p)})'$ for all j; (ii) For any l-dimensional vector $v = (v_1, ..., v_l)'$, the first component of v is denoted by $v^{(1)}$, and the subvector comprising its remaining components is denoted by $\tilde{v}$; (iii) $\alpha = (\alpha_1, \alpha_2)'$; (iv) For generic random vectors ξ_{jt} and ξ_{js}, $\xi_{j(ts)} \equiv \xi_{jt} - \xi_{js}$, e.g., $x_{1(12)} = x_{11} - x_{12} = (x_{1(12)}^{(1)}, ..., x_{1(12)}^{(p)})'$; (v) $\beta_0 \in \mathcal{B}$, the parameter space; and (vi) $\rho(b) = y_{1(12)} sgn(x_{1(12)}' b)$ for all $b \in \mathcal{B}$.

Next, we outline the regularity conditions for point identification and consistency of our semiparametric estimator based on the objective function (3.2).

SP1 $\{(y_i, x_i)\}_{i=1}^n$ is a random sample of n observations, where $y_i \equiv (y_{i1}', y_{i2}')'$ and $x_i \equiv (x_{i1}', x_{i2}')'$.

SP2 $\mathcal{B} = \{b \in \mathbb{R}^p : |b_1| = 1\} \cap \Xi$, where Ξ is a compact subset of $\mathbb{R}^p$.

SP3 For almost all (x, α), (i) $\epsilon_t \overset{d}{=} \epsilon_s | \alpha, x$ for all $t \neq s$, (ii) $\epsilon_t | \alpha, x$ has absolutely continuous distribution on $\mathbb{R}^2$.

SP4 Without loss of generality, $x_{1(12)}^{(1)}$ has everywhere positive Lebesgue density conditional on $\tilde{x}_{1(12)}$ and conditional on $x_{2(12)}$ in a neighborhood of $x_{2(12)}$ near zero. The coefficient β_{01} on $x_{jt}^{(1)}$ is nonzero and normalized to have absolute value 1.

SP5 The support of $x_{1(12)}$ conditional on $x_{2(12)}$ in a neighborhood of $x_{2(12)}$ near zero is not contained in any proper linear subspace of $\mathbb{R}^p$.

SP6 $x_{2(12)} \in \mathbb{R}^p$ is absolutely continuously distributed with PDF $f(\cdot)$ that is bounded from above on its support and strictly positive in a neighborhood of zero.[7]

SP7 For all $b \in \mathcal{B}$, $f(\cdot)$ and $E[\rho(b)|x_{2(12)} = \cdot]$ are continuously differentiable on their support with bounded first-order derivatives.

SP8 $K : \mathbb{R}^p \mapsto \mathbb{R}$ is a density function of bounded variation that satisfies: (i) $\sup_{v \in \mathbb{R}^p} |K(v)| < \infty$, (ii) $\int K(v) dv = 1$, and (iii) $\int |v_l| K(v) dv < \infty$ for all $l \in \{1, ..., p\}$.

SP9 h_n is a sequence of positive numbers that satisfies: (i) $h_n \to 0$ as $n \to \infty$, and (ii) $n h_n^p / \log n \to \infty$ as $n \to \infty$.

The above conditions suffice for point identification and consistency of our proposed estimator as stated in the following theorem, which is proved in Section A.

[7] Without the absolute continuity assumption, the point identification and consistency results are still valid. This assumption is made here is only for easing the exposition in the proof.

Theorem 3.1. *β_0 is point identified relative to all $b \in \mathcal{B}\backslash\{\beta_0\}$. Let $\widehat{\beta}$ be a sequence of the solutions to the problem*

$$\max_{b\in\mathcal{B}} \sum_{i=1}^{n} K(x_{i2(12)}/h_n)\rho_i(b)$$

Then, $\widehat{\beta} \xrightarrow{p} \beta_0$.

Next, to examine the effect of dimensionality in the number of choices, we consider the case with $T = 2$ and $J + 1$ alternatives (numbered from 0 to J, $J \geq 2$). For notation convenience, denote $z_1 = (x'_{2(12)}, ..., x'_{J(12)})'$, $z_2 = y_{1(12)}$, and $z_3 = x_{1(12)}$. Accordingly, the objective function is written as

$$\sum_{i=1}^{n} K(z_{1i}/h_n)z_{2i}sgn(z'_{3i}b)$$

Assumptions SP6′ - SP9′ stated below strengthen regularity conditions on the existence and finiteness of moments higher than those required for consistency and assume additional smoothness to allow convergence at a faster rate.

SP6′ $z_1 \in \mathbb{R}^{(J-1)p}$ is absolutely continuously distributed with bounded density $f_{z_1}(\cdot)$. Both $f_{z_1}(\cdot)$ and the conditional density $f_{z_1|z_3,z_2\neq 0}(\cdot)$ are strictly positive in a neighborhood of zero.

SP7′ For all $b \in \mathcal{B}$, $f_{z_1}(\cdot)$ and $E[\rho(b)|z_1 = \cdot]$ are twice differentiable on their support with bounded second-order derivatives where $\rho(b) = (y_{11} - y_{12})sgn\big((x_{11} - x_{12})'b\big)$.

SP8′ $K : \mathbb{R}^{(J-1)p} \mapsto \mathbb{R}$ is a density function of bounded variation that satisfies: (i) $\sup_{v\in\mathbb{R}^p} |K(v)| < \infty$, (ii) $\int K(v)dv = 1$, (iii) $\int vK(v)dv = 0$, and (iv) $\int |v_l|K(v)dv < \infty$ and $\int v_l^2 K(v)dv < \infty$ for all $l \in \{1, ..., (J-1)p\}$. [K is with bounded support]

SP9′ h_n is a sequence of positive numbers such that as $n \to \infty$: (i) $h_n \to 0$, (ii) $nh_n^{(J-1)p}/\log n \to \infty$, and (iii) $nh_n^{(J-1)p+3} \to 0$.

Under these conditions, the following theorem establishes the rate of convergence of the proposed estimator as a function of the number of choices J.

Theorem 3.2. *Let Assumptions SP1 - SP5 and SP6′ - SP9′ hold and $\widehat{\beta}$ be a sequence of the solutions to the problem*

$$\max_{b\in\mathcal{B}} \sum_{i=1}^{n} K(z_{1i}/h_n)z_{2i}sgn(z'_{3i}b)$$

Then, $|\widehat{\beta} - \beta_0|_2 = O_p((nh_n^{(J-1)p})^{-1/3})$, where $|\cdot|_2$ denotes the l_2 (Euclidean) norm.

We note that here, in contrast to cross-sectional case there are not "enough" matches for standard asymptotics to hold. In addition and more interestingly, in the multinomial panel data setting, rates of convergence depend on the number of choices, J as with more alternatives, we are matching more covariates. Proofs of the above results are collected in Section A.

3.2 Dynamic Multinomial Choice

We extend the base model of the previous section by examining the question of inference in a *dynamic* version of the multinomial panel data model. We follow the literature here and focus our inference problem on finite dimensional coefficient vectors which include in this section, the coefficient on the lagged choice variable.

In many situations, such as in the study of labor force and union participation, transportation choice, or health insurance carrier, it is observed that an individual who has experienced an event, or made some choice in the past is more likely to experience the event or make that same choice in the future as compared to another individual who has not experienced the event or made that choice. Heckman (1981) and Heckman (1991) discuss two explanations for this phenomenon. The first explanation is the presence of "true state dependence" in the sense that the lagged choice/decision enters the model as an explanatory variable and so having experience the event in the past, an economic agent is more likely to experience it in the future (due to familiarity for example). The second explanation that is advanced to explain this empirical regularity is the presence of serial correlation in the unobserved transitory errors that are in the model and this explanation revolves around heterogeneity (rather than state dependence): some individuals are more likely to make a given choice than other due to unobserved factors. The econometrics literature on the topic has provided various models to try and disentangle these two explanations. We contribute to this literature.

In particular, we expand results from the previous section by presenting identification and estimation methods for discrete choice models with structural state dependence that allow for the presence of unobservable individual heterogeneity in panels with a large number of individuals observed over a small number of time periods. Our results focus on point identification. We illustrate the approach with $J = 3$. A particular model that we consider can be expressed here as follows.

$$
\begin{aligned}
y^*_{i0t} &= 0 \\
y^*_{i1t} &= x'_{i1t}\beta_0 + \gamma_0 \mathbf{1}[y_{i1(t-1)} = 1] + \alpha_{i1} - \epsilon_{i1t} \\
y^*_{i2t} &= x'_{i2t}\beta_0 + \alpha_{i2} - \epsilon_{i1t}
\end{aligned}
$$

In this model, the parameters of interest are β_0 and γ_0. Identification is more complicated in dynamic models, even for binary choice. For example, Chamberlain (1985) shows that β_0 is *not* identified when there are 3 time periods, $t = 0, 1, 2$.[8] Honoré and Kyriazidou (2000) show point identification[9] of β_0 and γ_0 when there are 4 time periods, $t = 0, 1, 2, 3$. Their identification is based on conditioning on the subset of the population whose regressors do not change in periods 2 and 3. Finally, Khan et al. (2019) derive sharp bounds for coefficients in dynamic binary choice models with fixed effects under weak conditions (allowing for time trends, time dummies, etc).

Our identification strategy for the dynamic multinomial choice model is based on conditioning on the subpopulation whose regressors are time-invariant in different manners, depending on which alternative they are associated with. Specifically, in the three choices, four time periods setting above we condition on the subpopulation whose regressor values for choice 2 do not change in period 1, 2 and 3 and whose regressor values for choice 1 do not change over time in period 2 and 3.

After such conditioning, the problem reduces to identifying parameters in a dynamic binary choice model, for which existing methods can be applied. For example, if the post conditioning model is a dynamic Logit, which would arise if we begin with a dynamic multinomial Logit, we can use the method proposed in Honoré and Kyriazidou (2000), which is valid for four time periods. An attractive feature of their procedure is that when the covariates are discrete, the estimator will converge at the parametric rate with a limiting normal distribution, so conducting inference is relatively easy. We demonstrate both methods for the dynamic multinomial choice model considered here.

For the dynamic multinomial Logit model, we use the following conditional likeli-

[8] But γ_0 is identified if $\beta_0 = 0$.

[9] Their point identification result requires further restrictions on the serial behavior of the exogenous regressors that rules out, among other things, time trends as regressors. Our identification result for the dynamic multinomial choice imposes similar restrictions and so also does not allow for time trends as regressors.

hood function:

$$G_n^{DP,Logit}(b,g) = \sum_{i=1}^{n} \mathbf{1}[x_{i21} = x_{i22} = x_{i23}, x_{i12} = x_{i13}]\mathbf{1}[y_{i11} \neq y_{i12}] \times \log\left(\frac{\exp\left((x_{i11} - x_{i12})'b + g(d_{i0} - d_{i3})\right)^{y_{i11}}}{1 + \exp\left((x_{i11} - x_{i12})'b + g(d_{i0} - d_{i3})\right)}\right)$$

where $d_{i\cdot} \in \{0,1\}$. Note that scale normalization is no longer needed for maximum likelihood estimation. Honoré and Kyriazidou (2000) propose a multinomial Logit estimator whose identification and estimation are based on sequences of choices where the individual switches between alternatives at least once during the periods 1 through 3. For general J and T, the number of such sequences is $J^T - J^2$, then coding the estimator may be cumbersome, especially for cases with large J and/or large T.[10] Our estimator differs from theirs, as here we effectively transform a multinomial choice problem to a binary choice problem through matching x_{21} and x_{22}, which makes it considerably easier to implement.

We note here that in the case when all the regressors across all choices are discretely distributed, the estimator can be shown to converge at the parametric rate with a limiting normal distribution, as was shown in Honoré and Kyriazidou (2000) for the binary model.

For the semiparametric model, the objective function is of the form

$$G_n^{DP}(b,g) = \sum_{i=1}^{n} \mathbf{1}[x_{i21} = x_{i22} = x_{i23}, x_{i12} = x_{i13}](y_{i11} - y_{i12})sgn\left((x_{i11} - x_{i12})'b + g(y_{i13} - y_{i10})\right) \quad (3.3)$$

Note that for point identification, in this case, we require that one of the components of the regressors for the first choice has to be continuously distributed. Consequently, when matching regressors for this choice, we would need to assign kernel weights as illustrated before.

Under the standard "initial conditions" assumption as in e.g., Honoré and Kyriazidou (2000),[11] the maximizer of this objective function can be shown to be consistent, although as in the static model, the limiting distribution is nonstandard. Here we state the regularity conditions we impose to establish consistency. The proof is in Section A. To facilitate exposition of our conditions, we first introduce additional notation:

To simplify exposition, we first introduce extra notations for the dynamic panel setting: (i) $\theta_0 \equiv (\beta_0', \gamma_0)' \in \Theta$, the parameter space; (ii) $\psi(\theta) \equiv y_{1(12)}sgn(x_{1(12)}'b + gy_{1(03)})$

[10] For example, in our empirical illustration, we have $J = 4$ and $T_{max} = 77$ (unbalanced panel).

[11] Specifically, for the model at hand, the initial conditions assumption would be that $P(y_{i0} = 1)$ depends on $x_{i11}, x_{i12}, x_{i13}, \alpha_{i1}$.

for all $\theta \in \Theta$; and (iii) $\Delta x \equiv (x'_{2(12)}, x'_{2(23)}, x'_{1(23)})'$ and the event $\Omega \equiv \{\Delta x = 0\}$. Here we deliberately keep the notation as close as possible to Honoré and Kyriazidou (2000). Then, we outline the regularity conditions for point identification and consistency of our semiparametric estimator based on the objective function (3.3).

DP1 $\{(y_i, x_i)\}_{i=1}^n$ is a random sample of n observations, where $y_i \equiv (y'_{i0}, y'_{i1}, y'_{i2}, y'_{i3})'$ and $x_i \equiv (x'_{i1}, x'_{i2}, x'_{i3})'$.

DP2 $\Theta = \{\theta = (b', g)' \in \mathbb{R}^{p+1} : |b_1| = 1\} \cap \Xi$, where Ξ is a compact subset of $\mathbb{R}^{p+1}$.

DP3 For almost all (x, α), (i) $\epsilon_t \perp (x, y_0)|\alpha$ holds for all $t = 1, 2, 3$, (ii) $\epsilon_t|\alpha$ is iid over time[12] having absolutely continuous distribution on $\mathbb{R}^2$, and (iii) $P(y_{10} = 1|x, \alpha, \Omega) \in (0, 1)$.

DP4 Without loss of generality, $x^{(1)}_{1(12)}$ has everywhere positive Lebesgue density conditional on $\tilde{x}_{1(12)}$ and conditional on Δx in a neighborhood of Δx near zero. The coefficient β_{01} on $x^{(1)}_{jt}$ is nonzero and normalized to have absolute value 1.

DP5 The support of $x_{1(12)}$ conditional on Δx in a neighborhood of Δx near zero is not contained in any proper linear subspace of $\mathbb{R}^p$.

DP6 $\Delta x \in \mathbb{R}^{3p}$ is absolutely continuously distributed with PDF $f(\cdot)$ that is bounded from above on its support and strictly positive in a neighborhood of zero.[13]

DP7 For all $\theta \in \Theta$, $f(\cdot)$ and $E[\psi(\theta)|x_{2(12)} = x_{2(23)} = x_{1(23)} = \cdot]$ are continuously differentiable on their support with bounded first-order derivatives.

DP8 $K : \mathbb{R}^{3p} \mapsto \mathbb{R}$ is a density function of bounded variation that satisfies: (i) $\sup_{v \in \mathbb{R}^{3p}} |K(v)| < \infty$, (ii) $\int K(v)dv = 1$, and (iii) $\int |v_l|K(v)dv < \infty$ for all $l \in \{1, ..., 3p\}$.

DP9 h_n is a sequence of positive numbers that satisfies: (i) $h_n \to 0$ as $n \to \infty$, and (ii) $nh_n^{3p}/\log n \to \infty$ as $n \to \infty$.

The above conditions suffice for point identification and consistency of our proposed estimator as stated in the following theorem, also proved in Section A.

Theorem 3.3. *θ_0 is point identified relative to all $\theta \in \Theta \setminus \{\theta_0\}$. Let $\widehat{\theta} = (\widehat{\beta}', \widehat{\gamma}')'$ be a sequence of the solutions to the problem*

$$\max_{\theta \in \Theta} \sum_{i=1}^n K(\Delta x_i/h_n)\psi_i(\theta) = \max_{\theta \in \Theta} \sum_{i=1}^n K(\Delta x_i/h_n)(y_{i11} - y_{i12})sgn(x'_{i1(12)}b + gy_{i1(03)}))$$

Then, $\widehat{\theta} \xrightarrow{p} \theta_0$.

[12] Note that DP3 (ii) allows the distribution of $\epsilon_t|\alpha$ to vary across individuals.

[13] Without the absolute continuity assumption, the point identification and consistency results stated in Theorem 3.2 are still valid. This assumption made here is only for easing the exposition.

4 Simulation Study

In this section, we explore the relative finite sample performances of the proposed estimation procedures in cross-sectional and panel data (both static and dynamic) designs. We generate 1000 replications for six designs, using sample sizes ranging from 250 to 10000. In all designs, the regressor vector always has one and only one component that is continuously distributed (standard normal) with all the rest being binary, and the error vector for each individual follows a multivariate normal distribution that allows for correlation across components. Recall that to implement proposed estimators one must choose a kernel as well as a bandwidth for matching the continuous regressor. All the results presented in this section use a normal kernel and Silverman's rule of thumb to choose bandwidth.

For the cross-sectional model, we generate data from three designs, varying the number of regressors and/or the size of the choice set. The first two are for a model with three choices, and we increase the number of regressors from 3 to 5. This is meant to give an idea on the sensitivity of our estimator to the dimensionality of the regressor space. In the third design, we considered three regressors but five choices. Here, we aim to explore the sensitivity of our procedure to the dimensionality of the choice space.

For the panel data model, we generated data from two designs. The first is for a static panel data model with three choices and three regressors with two periods of data. The second panel data design is for the dynamic model where there are three choices and three regressors with the second of the two binary regressors being the lagged choice. For this model, we simulate four periods of data as this is the minimum T required for our point identification result.

For each of these six designs and varying sample sizes, we report the mean bias and root mean squared error (RMSE) of the corresponding estimator. Since these statistics can be sensitive to outliers, we also present the median bias and the median absolute deviation (MAD). Below we state the details of the designs considered and the summary statistics for our estimators.[14]

Our benchmark design (Design 1) for the cross-sectional model is based on the data generating process (DGP) with choice set $\{0, 1, 2\}$ and latent utility functions:

[14]For each of the panel data models, we also report results for the two-step estimator, where we construct the second step estimator based on matching index, computed based on the first step estimators.

$$y_{i0}^* = 0$$
$$y_{ij}^* = x_{ij}^{(1)} + \beta_1 x_{ij}^{(2)} + \beta_2 x_{ij}^{(3)} - \epsilon_{ij},\ j = 1, 2$$

where $x_{ij}^{(1)}, x_{ij}^{(2)}, x_{ij}^{(3)}$ denote the 3 components of the vector x_{ij}, $\beta_1 = \beta_2 = 1$, $x_{i1}^{(1)} \overset{iid}{\sim} N(0,1)$, $x_{i2}^{(1)} \overset{iid}{\sim}$ Bino$(1, 0.5)$, $x_{ij}^{(k)} \overset{iid}{\sim}$ Bino$(1, 0.5)$ for all $j \in \{1, 2\}$ and $k \in \{2, 3\}$, and

$$(\epsilon_{i1}, \epsilon_{i2}) \overset{iid}{\sim} \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

Table 1 reports the results for this benchmark design.

Table 1: (Design 1) 3 Choices, 3 Regressors, 2 Parameters

	β_1				β_2			
	Mean	RMSE	Median	MAD	Mean	RMSE	Median	MAD
$N = 250$	0.0161	0.4706	0.0104	0.3430	0.0182	0.4798	-0.0135	0.3383
$N = 500$	0.0418	0.3726	0.0190	0.2297	0.0428	0.3684	0.0224	0.2222
$N = 1000$	0.0138	0.2619	0.0022	0.1562	0.0098	0.2577	-0.0022	0.1585

As our cross-sectional estimator is "localized" (matching covariates associated with $J - 1$ alternatives), one may worried about that the dimensionality of the design (both in the regressor space and choice space) may have a large effect on the results in Monte Carlo studies. In order to investigate the finite sample performance of the proposed estimator in higher dimensional, more complicated designs, we consider the following two modifications of the benchmark design:

- Design 2: We keep the choice set and error distribution unchanged, while add two regressors to the benchmark design. Specifically, we consider the DGP with latent utility functions:

  $$y_{i0}^* = 0$$
  $$y_{ij}^* = x_{ij}^{(1)} + \beta_1 x_{ij}^{(2)} + \beta_2 x_{ij}^{(3)} + \beta_3 x_{ij}^{(4)} + \beta_4 x_{ij}^{(5)} - \epsilon_{ij},\ j = 1, 2$$

  where $\beta_1 = \beta_2 = 1$, $\beta_3 = \beta_4 = 0$, $x_{i1}^{(1)} \overset{iid}{\sim} N(0,1)$, and all other regressors are iid Bino$(1, 0.5)$. Note that the DGP is the same as for the benchmark case and the only difference is that two additional regressors are included in the estimation.

- Design 3: We keep the latent utility functions the same, while enlarge the choice set to be $\{0, 1, 2, 3, 4\}$, i.e., we consider the design with

$$y_{i0}^* = 0$$
$$y_{ij}^* = x_{ij}^{(1)} + \beta_1 x_{ij}^{(2)} + \beta_2 x_{ij}^{(3)} - \epsilon_{ij},\ j = 1, 2, 3, 4$$

where $\beta_1 = \beta_2 = 1$, $x_{i1}^{(1)} \overset{iid}{\sim} N(0, 1)$, all other regressors are iid Bino$(1, 0.5)$, and

$$(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}, \epsilon_{i4}) \overset{iid}{\sim} \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}\right)$$

The results of these two experiments are summarized in Table 2[15] and 3, respectively.

Table 2: (Design 2) 3 Choices, 5 Regressors, 4 Parameters

	β_1				β_2			
	Mean	RMSE	Median	MAD	Mean	RMSE	Median	MAD
$N = 250$	0.0735	0.5245	0.0631	0.3830	0.0313	0.5453	0.0099	0.4151
$N = 500$	0.0551	0.4147	0.0259	0.2715	0.0824	0.4144	0.0727	0.2823
$N = 1000$	0.0294	0.3014	0.0033	0.2089	0.0302	0.2915	0.0039	0.1907

Table 3: (Design 3) 5 Choices, 3 Regressors, 2 Parameters

	β_1				β_2			
	Mean	RMSE	Median	MAD	Mean	RMSE	Median	MAD
$N = 250$	-0.1279	0.6581	-0.1686	0.5775	-0.0879	0.6473	-0.1160	0.5805
$N = 500$	-0.0955	0.6196	-0.1180	0.5200	-0.0459	0.6102	-0.0719	0.5025
$N = 1000$	-0.0372	0.5724	-0.0375	0.4604	-0.0334	0.5736	-0.0532	0.4670
$N = 2000$	0.0122	0.5099	-0.0057	0.3786	0.0132	0.4979	-0.0142	0.3601

As our results demonstrate, the performance is in line with the asymptotic theory. Specifically, the cross-sectional estimator is root-n consistent as both the bias and RMSE

[15] To conserve space, we report only the results for β_1 and β_2.

shrink at the parametric rate. This seems true regardless of the number of regressors, though as expected performance for each sample size deteriorates with the number of regressors. However, that is not the case as we increase the size of the choice set. As seen in Table 3, with five choices, the finite sample performance is relatively poor, and furthermore, does not improve with larger sample sizes as well as it did in the other designs. Thus it appears to us that for this model the adversarial effects of dimensionality lie in the choice dimension and not as much in the regressor dimension.[16]

We then turn to examine the finite sample properties of the maximum score estimators for panel data multinomial choice models. We start from the static panel case and consider the design (Design 4) with choice set $\{0, 1, 2\}$ and a panel of two time period ($T = 2$). The latent utility functions for individual i in time period $t \in \{1, 2\}$ are

$$\begin{aligned} y^*_{i0t} &= 0 \\ y^*_{ijt} &= x^{(1)}_{ijt} + \beta_1 x^{(2)}_{ijt} + \beta_2 x^{(3)}_{ijt} + \alpha_{ij} - \epsilon_{ijt},\ j = 1, 2 \end{aligned}$$

where $\beta_1 = \beta_2 = 1$, $x^{(1)}_{i1t} \overset{iid}{\sim} N(0, 1)$ for all t, all other regressors are iid Bino$(1, 0.5)$, and

$$(\epsilon_{i11}, \epsilon_{i21}, \epsilon_{i12}, \epsilon_{i22}) \overset{iid}{\sim} \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}\right)$$

The fixed effects are generated as $\alpha_{i1} = T^{-1} \sum_{t=1}^{T} x_{i1t}$ and $\alpha_{i2} = T^{-1} \sum_{t=1}^{T} x_{i2t} - 0.5$. In Table 4 and 5, we report respectively the results for this static panel design using one-step and two-step maximum score estimators.

Table 4: (Design 4) 3 Choices, 3 Regressors, 2 Parameters, 2 Periods

	β_1				β_2			
	Mean	RMSE	Median	MAD	Mean	RMSE	Median	MAD
$N = 500$	-0.0570	0.6118	-0.0918	0.5112	-0.0716	0.6110	-0.0661	0.5247
$N = 1000$	-0.0468	0.5860	-0.0647	0.4819	-0.0524	0.5843	-0.0650	0.4802
$N = 2000$	-0.0401	0.5673	-0.0544	0.4399	-0.0417	0.5587	-0.0644	0.4366
$N = 5000$	-0.0221	0.4932	-0.0476	0.3555	0.0044	0.4911	-0.0043	0.3682
$N = 10000$	0.0006	0.4530	-0.0109	0.3182	0.0084	0.4510	-0.0048	0.3235

[16] It is not too surprising that our localized estimator performs relatively better in Design 2 than in Design 3. To implement the proposed estimator, one need match $(J - 2) \times p$ regressors, where p is the number of regressors associated with each alternative. This number for Design 2 is 5, while for Design 3 it is 9.

Table 5: (Design 4, Two-step) 3 Choices, 3 Regressors, 2 Parameters, 2 Periods

	β_1				β_2			
	Mean	RMSE	Median	MAD	Mean	RMSE	Median	MAD
$N = 500$	-0.0539	0.6008	-0.0580	0.5144	-0.0562	0.5895	-0.0513	0.4842
$N = 1000$	-0.0413	0.5978	-0.0479	0.5134	-0.0497	0.5732	-0.0477	0.4717
$N = 2000$	-0.0252	0.5557	-0.0356	0.4311	0.0154	0.5632	0.0014	0.4465
$N = 5000$	0.0329	0.4930	-0.0033	0.3598	0.0017	0.4928	-0.0203	0.3568
$N = 10000$	0.0256	0.4438	0.0065	0.3149	0.0389	0.4415	0.0100	0.3131

Our dynamic panel design (Design 5) has the same choice set as the static design but four time periods ($T = 3$, $t \in \{0, 1, 2, 3\}$). The latent utility functions are

$$\begin{aligned}
y^*_{i0t} &= 0,\ t = 0, 1, 2, 3 \\
y^*_{ij0} &= x^{(1)}_{ij0} + \beta x^{(2)}_{ij0} + \alpha_{ij} - \epsilon_{ij0},\ j = 1, 2 \\
y^*_{i1t} &= x^{(1)}_{i1t} + \beta x^{(2)}_{i1t} + \gamma y_{i1(t-1)} + \alpha_{i1} - \epsilon_{i1t},\ t = 1, 2, 3 \\
y^*_{i2t} &= x^{(1)}_{i2t} + \beta x^{(2)}_{i2t} + \alpha_{i2} - \epsilon_{i2t},\ t = 1, 2, 3
\end{aligned}$$

where $(\beta, \gamma) = (1, 0.5)$, $y_{i1(t-1)} = \mathbf{1}[u_{i1(t-1)} > \max\{0, u_{i2(t-1)}\}]$, $x^{(1)}_{i1t} \overset{iid}{\sim} N(0, 1)$ for all t, all other regressors are iid Bino$(1, 0.5)$, and

$$(\epsilon_{i1t}, \epsilon_{i2t}) \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

independent across i and over time. We use the same way to generate the fixed effects as the static design. One-step and two-step estimation results for this design are summarized in Table 6 and 7, respectively.

For the panel data results, the static panel data estimator also appears to be consistent but appears to converge more slowly in terms of bias and RMSE. It takes samples sizes that are larger than 2000 before the estimator performs adequately well. For the semi-parametric dynamic panel data model, results seem worse still and appear to improve even more slowly with increases in the sample size. In both cases, the two-step estimator improves finite sample performance a little, particularly in the dynamic model.

Table 6: (Design 5) 3 Choices, 3 Regressors, 2 Parameters, 4 Periods

	β				γ			
	Mean	RMSE	Median	MAD	Mean	RMSE	Median	MAD
$N = 500$	0.0005	0.4706	0.0138	0.3331	-0.0281	0.5356	-0.0265	0.4465
$N = 1000$	0.0196	0.4850	0.0163	0.3551	-0.0491	0.5241	-0.0708	0.4160
$N = 2000$	0.0000	0.4685	0.0175	0.3616	-0.0610	0.5380	-0.0762	0.4194
$N = 5000$	-0.0190	0.4389	-0.0047	0.2614	-0.0348	0.5254	-0.0611	0.4289
$N = 10000$	-0.0226	0.4246	-0.0047	0.2451	-0.0665	0.5176	-0.0898	0.3846

Table 7: (Design 5, Two-step) 3 Choices, 3 Regressors, 2 Parameters, 4 Periods

	β				γ			
	Mean	RMSE	Median	MAD	Mean	RMSE	Median	MAD
$N = 500$	0.0151	0.4786	0.0161	0.3242	-0.0290	0.5392	-0.0304	0.4356
$N = 1000$	0.0077	0.4571	0.0113	0.3158	-0.0665	0.5649	-0.0801	0.4641
$N = 2000$	0.0150	0.4583	0.0047	0.2914	-0.0586	0.5250	-0.0816	0.4092
$N = 5000$	0.0047	0.4218	0.0049	0.2488	-0.0538	0.5009	-0.0823	0.3597
$N = 10000$	0.0012	0.3959	-0.0005	0.2158	-0.0625	0.4795	-0.0921	0.3530

As a final component of our simulation study, we explore how the conditional Logit estimator performs in the dynamic panel design. As the point identification of Logit does not rely on the existence of a continuous regressor, we let all regressors in Design 5 be iid with Bernoulli distribution. It is easy to show that the Logit estimator would be root-n consistent[17] if the conditional likelihood function is correctly specified. However, with errors not satisfying the IIA property, we would expect the Logit estimator to be inconsistent. The simulation here aims to explore the sensitivity of the parametric estimator to model misspecification. In Table 8 for Logit results, inconsistency is clearly demonstrated with biases exceeding 50% even at sample sizes of 10000. The inconsistency is to

[17]Note that if we have one continuous regressor and use bandwidth $h_n = O(n^{-1/5})$, the Logit estimator is asymptotically biased even if the model is correctly specified.

be expected as the logit estimator is based on iid type 1 extreme value errors.

Table 8: (Design 5, Logit) 3 Choices, 3 Regressors, 2 Parameters, 4 Periods

	β				γ			
	Mean	RMSE	Median	MAD	Mean	RMSE	Median	MAD
$N = 500$	0.6760	0.8963	0.6036	0.3433	1.5346	3.6605	0.4860	2.0067
$N = 1000$	0.6535	0.7731	0.6337	0.2569	1.8503	3.8376	2.0596	2.3813
$N = 2000$	0.5875	0.6585	0.5627	0.1903	1.8067	3.5593	1.4309	2.3469
$N = 5000$	0.5863	0.6171	0.5799	0.1302	1.6068	2.9385	0.8063	1.3635
$N = 10000$	0.5858	0.6052	0.5844	0.0994	1.0708	2.1002	0.5929	0.8105

5 Empirical Illustration

In this section we also illustrate the finite sample properties of our new rank estimator by applying it to the often used optical scanner panel data set on purchases of saltine crackers in the Rome (Georgia) market, that was collected by Information Resources Incorporated. The data set contains information on all purchases of crackers (3292) of 136 households over a period of two years, including brand choice, actual price of the purchased brand and shelf price of other brands, and whether there was a display and/or newspaper feature of the considered brands at the time of purchase. A subset of this data set was analyzed in Jain, Vilcassim, and Chintagunta (1994) as well as Paap and Frances (2000).

Table 9: Data Characteristics of Saltine Crackers

	Sunshine	Keebler	Nabisco	Private
Market Share	0.07	0.07	0.54	0.32
Display	0.13	0.11	0.34	0.10
Feature	0.04	0.04	0.09	0.05
Average Price	0.96	1.13	1.08	0.68

Table 9 summarizes some data characteristics of saltine crackers. There are three major national brands in the database: Sunshine, Keebler, and Nabisco, with market shares of

7%, 7%, and 54%, respectively. Local brands are aggregated and referred to in the table as "Private" label, which has a market share of 32%. The data set also includes three explanatory variables, two of which are binary and one of which is continuous. The first binary explanatory variable, which we will refer to as "display", denotes whether or not a brand was on special display at the store at the time of purchase. The second binary explanatory variable, which we will refer to as "feature", denotes whether or not a brand was featured in a newspaper advertisement at the time of purchase. Table 9 reports fractions for the binary variables, so for example, the numbers in the "display" row correspond to fractions of purchase occasions on which each brand is on display. The third explanatory variable we will use is the "price" which corresponds to the price of a brand. This explanatory variable has rich enough support in the data set that we feel that treating it as a continuously distributed random variable is a reasonable approximation. Table 9 reports the sample average of the price of each brand over the 3292 purchases.

There are two features of this data set that make it particularly suitable to apply our semiparametric procedures. One is that there is one continuous regressor (price) which is needed for point identification. Importantly, the other regressors are binary, so the "matching" part of our procedure can be implemented relatively easily. The second important feature is that the data is actually a panel data set based on 136 households making purchase decisions over a period of two years. Thus we can use this data to apply both our cross-sectional (pooled) estimator as well as panel data (static and dynamic) estimators.

Specifically, for this data set here, we apply our rank estimators to the multinomial choice model with four choices and three regressors. As mentioned above, existing work such as Jain et al. (1994) and Paap and Frances (2000) have used this data to estimate *parametric* multinomial choice models. Thus our semiparametric approach would indicate how sensitive their results and conclusions are to the strong assumptions they imposed, either in the way of parametric assumptions such as multinomial probit specification, and/or ignoring the unobserved heterogeneity that our fixed effect estimators allow for in the panel data setting. This can be done by comparing the estimates we get for the regression coefficients using our methods to those attained in Paap and Frances (2000).

In what follows, we denote the choice set as $\mathcal{J} = \{1 = \text{Nabisco}, 2 = \text{Sunshine}, 3 = \text{Keebler}, 4 = \text{Private}\}$. The observed choice and explanatory variables are measured as follows: For each household i, brand j, and purchase t,

- y_{it}: observed choice ($= 1, 2, 3$, or 4).
- $x_{ijt}^{(1)}$: normalized "price" ($mean = 0$, $std = 1$).

- $x_{ijt}^{(2)}$: "display", 0-1 valued.
- $x_{ijt}^{(3)}$: "feature", 0-1valued.

Note that the data set is an unbalanced panel with $n = 136$ households and T varying with i ($\min\{T_i\} = 14$, $\max\{T_i\} = 77$).

Following Paap and Frances (2000), we model the latent utility of household i for brand j in the t-th purchase as

$$y_{ijt}^* = -x_{ijt}^{(1)} + \beta_1 x_{ijt}^{(2)} + \beta_2 x_{ijt}^{(3)} - \epsilon_{ijt}$$

where the coefficient on $x_{ijt}^{(1)}$ is normalized to be -1, and (β_1, β_2) are regression coefficients to be estimated. ϵ_{ijt} is the unobserved scalar disturbance term.

To implement our cross-sectional rank estimator, we pool the cross-section (i) and "time-series" (t) aspects of the panel. The estimation was implemented in R, using the differential evolution algorithm to attain a global optimum of the objective function with respect to the coefficients on "display" and "feature". To construct the objective function we matched on the continuous variable (price) using a normal kernel function and Silverman's rule of thumb to pick the bandwidth. The computation was relatively fast - the estimator of the two coefficients took only 3 minutes to attain using a MacBook Pro laptop.

To attain confidence regions we employed the standard bootstrap by sampling from the original data set (with replacement) and computing the rank estimator for each sampled data set. Employing this for 500 sampled data sets took about 25 hours. Point estimates and confidence regions for each of the two coefficients on the binary regressors, denoted respectively by β_1 and β_2, are reported in Table 10. For comparison purposes, the table also reports results from estimators for two parametric models, multinomial Probit and multinomial Logit. To compare parametric results to semiparametric ones, in the former case we report the ratio of coefficients of the binary regressors to the absolute value of the coefficient on "price".

Table 10: Parametric and Semiparametric Estimates for Cross Sectional Model

	β_1	95% CI of β_1	β_2	95 % CI of β_2
Semiparametric	0.3331	(0.1010, 0.4918)	0.3081	(0.1006, 0.4978)
Multinomial Logit	0.1368	(-0.0480, 0.3215)	0.7381	(0.4268, 1.0495)
Multinomial Probit	0.0919	(-0.0855, 0.2693)	0.6185	(0.3090, 0.9280)

As we can see, the results are strikingly different. For the parametric estimators for multinomial Probit relative coefficients for display and feature are 0.1226 and 0.9608, respectively. For multinomial Logit, they are 0.2150 and 1.1829. In each parametric setting, the coefficient (ratio) on display is not significantly different from 0 at the 95% level, whereas the coefficient on feature is significantly positive. For our semiparametric estimates, the results are coefficient estimates of $(0.3331, 0.3081)$. In contrast to the parametric results, each coefficient is significantly positive at the 95% level.

Now we turn attention to the panel data features of the data set. For the static model, we consider the following specification

$$y^*_{ijt} = -x^{(1)}_{ijt} + \beta_1 x^{(2)}_{ijt} + \beta_2 x^{(3)}_{ijt} + \alpha_{ij} - \epsilon_{ijt}$$

where α_{ij} collects the individual and choice specific effects.

Employing our estimator, our results were $(\hat{\beta}_1, \hat{\beta}_2) = (3.6924, 0.4472)$ with criterion function = 864.0031 . These results are interesting when compared to results attained using parametric and semiparametric estimators for the cross sectional model. In the panel data model the coefficient on display is much larger than the coefficient on feature. This is in complete contrast to multinomial Probit and Logit where the coefficient on display is not statistically different from 0 and the coefficient on feature is significantly positive. The panel data estimates are also different from the semiparametric estimates for the cross sectional model, where the coefficients on display and feature are virtually identical. However, it should be emphasized that for panel data estimates we only report point estimates and not confidence regions. This is because the limiting distribution theorem of either panel estimator has not been derived, and we conjecture distribution theory will be nonstandard so it is unlikely that the standard bootstrap can provide valid confidence regions in this setting.

Table 11: Parametric and Semiparametric Estimates for Static Panel Data Model

	β_1	β_2
Semiparametric	0.4489	0.4528
Conditional Logit	-0.0639	0.5838

For each i and $t \in \{2, ..., T_i\}$,

$$y^*_{i1t} = -x^{(1)}_{i1t} + \beta_1 x^{(2)}_{i1t} + \beta_2 x^{(3)}_{i1t} + \gamma y_{i1(t-1)} + \alpha_{i1} - \epsilon_{i1t}$$
$$y^*_{ijt} = -x^{(1)}_{ijt} + \beta_1 x^{(2)}_{ijt} + \beta_2 x^{(3)}_{ijt} - \epsilon_{ijt},\ j = 2, 3, 4$$

where as above $y_{i1t} = \mathbf{1}[y_{it} = 1]$.

Employing each of our two estimators for the dynamic model, our estimation results were $(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\gamma}) = (1.5041, 1.4408, 0.5710)$ with criterion function = 3.118645 for the semiparametric estimator, and $(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\gamma}) = (0.1274, 1.6865, 0.6185)$ for the Logit. Note for both the semiparametric and Logit estimates the first two estimated coefficients are very different when compared to the static model, indicating the dynamic specification may be relevant for this data set, and ignoring this aspect can lead to misspecification. This point is consistent with the estimated coefficient on lagged choice being quite different from zero, indicating "persistence" in consumer behavior for this product.

Table 12: Parametric and Semiparametric Estimates for Dynamic Panel Data Model

	β_1	β_2	γ
Semiparametric	0.6024	1.2716	0.4005
Conditional Logit	0.8270	2.2931	1.2091

6 Conclusions

In this paper we proposed new estimation procedures for semiparametric multinomial choice models. For the cross-sectional model we proposed a local rank based procedure, which was shown to be root-n consistent and asymptotically normal, even in designs where no smoothing parameters were required. The pairwise differencing is readily extended to time differencing, enabling a consistent estimator for a panel data estimator of a model with choice and individual specific effects. Furthermore we attain a new identification result for a dynamic multinomial choice model with lagged discrete dependent variables, and proposed new consistent estimators for the coefficients on coefficients on the lagged dependent variables.

The work here leaves many open areas for future research. For example limiting distribution theory needs to be established for the panel data estimators. Also, as pointed out, in both panel data settings the propose procedure suffers from a curse of dimensionality in the number of choices. It is thus an open question if our proposed approach result in a rate optimal estimator. Rate optimality for dynamic binary choice models was discussed in Seo and Otsu (2018), but such bounds are lacking in the multinomial case.

References

ABREVAYA, J., J. HAUSMAN, AND S. KHAN (2010): "Testing for Causal Effects in a Generalized Regression Model with Endogenous Regressors," *Econometrica*, 78, 2043–2061.

AHN, H., J. POWELL, H. ICHIMURA, AND P. RUUD (2017): "Simple Estimators for Invertible Index Models," *Journal of Business Economics and Statistics*, 36, 1–10.

AMEMIYA, T. (1985): *Advanced Econometrics*, Harvard University Press.

ANDERSEN, E. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society*, 32, 283–301.

ARELLANO, M. AND S. BONHOMME (2009): "Robust priors in nonlinear panel data models," *Econometrica*, 77, 489–536.

ARELLANO, M. AND B. HONORÉ (2001): "Panel Data Models: Some Recent Developments," *Handbook of econometrics.Volume 5*, 3229–96.

BONHOMME, S. (2012): "Functional Differencing," *Econometrica*, 80, 1337–1385.

CHAMBERLAIN, G. (1984): "Panel Data," in *Handbook of Econometrics, Vol. 2*, ed. by Z. Griliches and M. Intriligator, North Holland.

——— (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Logitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer, Cambridge University Press.

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78, 159–168.

CHEN, S., S. KHAN, AND X. TANG (2015): "Informational Content in Static and Dynamic Discrete Response Panel Data Models," U Penn Working Paper.

CHERNOZHUKOV, V., I. F. VAL, J. HAHN, AND W. NEWEY (2013): "Average and Quantile Effects in Non Separable Panel Data Models," *Econometrica*, 535–580.

DUBÉ, J.-P., G. J. HITSCH, AND P. E. ROSSI (2010): "State dependence and alternative explanations for consumer inertia," *The RAND Journal of Economics*, 41, 417–445.

GAO, W. AND M. LI (2019): "Robust Semiparametric Estimation in Panel Multinomial Choice Models," Working Paper.

GRAHAM, B. AND J. POWELL (2012): "Identification and Estimation of 'Irregular' Correlated Random Coefficient Models," *Econometrica*, 80, 2105–2152.

HAN, A. K. (1987): "Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 357–362.

HANDEL, B. R. (2013): "Adverse selection and inertia in health insurance markets: When nudging hurts," *American Economic Review*, 103, 2643–82.

HECKMAN, J. (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46, 931–960.

——— (1981): "Statistical Models for Discrete Panel Data," in *Structural Analysis of Discrete Data*, ed. by C. Manski and D. McFadden, MIT Press.

——— (1991): "Identifying the Hand of the Past: Distinguishing State Dependence from Heterogeneity," *American Economic Review*, 75–79.

HODERLEIN, S. AND H. WHITE (2012): "Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects," *Journal of Econometrics*, 168, 300–314.

HONORÉ, B. AND E. KYRIAZIDOU (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

ILLANES, G. (2016): "Switching Costs in Pension Plan Choice," *Unpublished manuscript*.

JAIN, D., N. VILCASSIM, AND P. CHINTAGUNTA (1994): "A Random-Coefficients Logit Brand-Choice Model Applied to Panel data," *Journal of Business Economics and Statistics*, 12, 317–328.

KETCHAM, J. D., C. LUCARELLI, AND C. A. POWERS (2015): "Paying attention or paying too much in Medicare Part D," *American Economic Review*, 105, 204–33.

KHAN, S., M. PONOMAREVA, AND E. TAMER (2016): "Identification in Panel Data Models with Endogenous Censoring," *Journal of Econometrics*, 194, 57–75.

——— (2019): "Identification of Dynamic Panel Binary Response Models," Working Paper.

KHAN, S. AND E. TAMER (2018): "Discussion of Simple Estimators for Invertible Index Models by H. Ahn, H. Ichimura, J. Powell, and P. Ruud," *Journal of Business & Economic Statistics*, 36, 11–15.

KIM, J. AND D. POLLARD (1990): "Cube root asymptotics," *The Annals of Statistics*, 18, 191–219.

KLEIN, R. AND R. SPADY (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

LEE, L.-F. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381–428.

MANSKI, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3(3), 205–228.

——— (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357–362.

MCFADDEN, D. (1978): "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Residential Location*, ed. by A. K. et. al., North Holland Pub. Co.

MERLO, A. AND K. WOLPIN (2015): "The Transition from School to Jail: Youth Crime and High School Completion Among Black Males," Working Paper.

NOLAN, D. AND D. POLLARD (1987): "U-Processes: Rates of Convergence," *Annals of Statistics*, 15, 780–799.

PAAP, R. AND P. H. FRANCES (2000): "A dynamic multinomial probit model for brand choice with different long-run and short-run effects of marketing-mix variables," *Journal of Applied Econometrics*, 15, 717–744.

PAKES, A. AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

PAKES, A. AND J. PORTER (2014): "Moment Inequalties for Multinomial Choice with Fixed Effects," Harvard University Working Paper.

POLYAKOVA, M. (2016): "Regulation of insurance with adverse selection and switching costs: Evidence from Medicare Part D," *American Economic Journal: Applied Economics*, 8, 165–95.

RAVAL, D. AND T. ROSENBAUM (2018): "Why Do Previous Choices Matter for Hospital Demand? Decomposing Switching Costs from Unobserved Preferences," *Review of Economics and Statistics*, forthcoming.

SEO, M. AND T. OTSU (2018): "Local M-estimation with discontinuous criterion for dependent and limited observations," *Annals of Statistics*, 46, 344–369.

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

SHI, X., M. SHUM, AND W. SONG (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

A Static and Dynamic Panel Data Estimators

A.1 Static Panel Data Estimators

A.1.1 Consistency

Let Ω denote the event $x_{2(12)} = 0$. To simplify notation, we define $z_1 = x_{2(12)}$, $z_2 = y_{1(12)}$, $z_3 = x_{1(12)}$,

$$Q(b) = f(0)E[\rho(b)|\Omega]$$

$$Q_n(b) = \frac{1}{nh_n^p}\sum_{i=1}^{n} K(z_{1i}/h_n)\rho_i(b)$$

and

$$\varphi(\cdot) = f(\cdot)E[\rho(b)|z_1 = \cdot]$$

In what follows, we focus on the case where $\mathcal{B} \subset \{b \in \mathbb{R}^p : b_1 = 1\}$. The case with $\mathcal{B} \subset \{b \in \mathbb{R}^p : b_1 = -1\}$ is symmetric.

Lemma A.1. *Under Assumptions SP3 - SP6, $Q(\beta_0) > Q(b)$ for all $b \in \mathcal{B} \setminus \{\beta_0\}$.*

Proof of Lemma A.1. Denote $\mathcal{Z}_b = \{z_3 : sgn(z_3'b) \neq sgn(z_3'\beta_0)\}$ for all $b \in \mathcal{B} \setminus \{\beta_0\}$. Note that $P(\tilde{z}_3'\tilde{b} = \tilde{z}_3'\tilde{\beta}_0|\Omega) < 1$ by Assumption SP5, and $P(z_3^{(1)} \in \mathcal{N}|\tilde{z}_3, \Omega) > 0$ by Assumption SP4, where $\mathcal{N} = \{-\tilde{z}_3'\tilde{b} < z_3^{(1)} < -\tilde{z}_3'\tilde{\beta}_0\} \cup \{-\tilde{z}_3'\tilde{\beta}_0 < z_3^{(1)} < -\tilde{z}_3'\tilde{b}\}$. Therefore,

$$P(\mathcal{Z}_b|\Omega) = P(z_3^{(1)} \in \mathcal{N}|\tilde{z}_3'\tilde{b} \neq \tilde{z}_3'\tilde{\beta}_0, \Omega)P(\tilde{z}_3'\tilde{b} \neq \tilde{z}_3'\tilde{\beta}_0|\Omega) > 0$$

Then, we have

$$\begin{aligned}
&Q(\beta_0) - Q(b)\\
=&f(0)E[z_2(sgn(z_3'\beta_0) - sgn(z_3'b))|\Omega]\\
=&2f(0)\int_{\mathcal{Z}_b} sgn(z_3'\beta_0)E[z_2|z_3, \Omega]dF_{z_3|\Omega}\\
=&2f(0)\int_{\mathcal{Z}_b} sgn(z_3'\beta_0)E[E[z_2|x, \alpha, \Omega]|z_3, \Omega]dF_{z_3|\Omega}\\
=&2f(0)\int_{\mathcal{Z}_b} E[sgn(z_3'\beta_0)(P(y_{11} = 1|x, \alpha, \Omega) - P(y_{12} = 1|x, \alpha, \Omega))|z_3, \Omega]dF_{z_3|\Omega}
\end{aligned}$$

Next, note that by definition,

$$P(y_{11} = 1|x, \alpha, \Omega) = P(x_{11}'\beta_0 + \alpha_1 - \epsilon_{11} > \max\{0, x_{21}'\beta_0 + \alpha_2 - \epsilon_{21}\}|x, \alpha, \Omega)$$

and

$$P(y_{12}=1|x,\alpha,\Omega)=P(x_{12}'\beta_0+\alpha_1-\epsilon_{12}>\max\{0,x_{22}'\beta_0+\alpha_2-\epsilon_{22}\}|x,\alpha,\Omega)$$

Hence, by Assumption SP3, we have $sgn(P(y_{11}=1|x,\alpha,\Omega)-P(y_{12}=1|x,\alpha,\Omega))=sgn(z_3'\beta_0)$. Furthermore, $P(y_{11}=1|x,\alpha,\Omega)=P(y_{12}=1|x,\alpha,\Omega)$ if and only if $z_3'\beta_0=0$ which is an event having zero probability measure under Assumption SP4. Then,

$$\begin{aligned}&E[sgn(z_3'\beta_0)(P(y_{11}=1|x,\alpha,\Omega)-P(y_{12}=1|x,\alpha,\Omega))|z_3,\Omega]\\=&E[|sgn(z_3'\beta_0)(P(y_{11}=1|x,\alpha,\Omega)-P(y_{12}=1|x,\alpha,\Omega))||z_3,\Omega]\\=&E[|P(y_{11}=1|x,\alpha,\Omega)-P(y_{12}=1|x,\alpha,\Omega)||z_3,\Omega]>0\end{aligned}$$

and the desired result follows from Assumption SP6. □

Proof of Theorem 3.1. The proof proceeds by verifying the four sufficient conditions for Theorem 9.6.1 in Amemiya (1985): (C1) $\mathcal{B}$ is a compact set, (C2) $Q_n(b)$ is a measurable function for all $b\in\mathcal{B}$, (C3) $Q_n(b)$ converges in probability to a nonstochastic function $Q(b)$ uniformly for all $b\in\mathcal{B}$, (C4) $Q(b)$ is continuous in b and is uniquely maximized at β_0.

The compactness of $\mathcal{B}$ is satisfied by construction. Condition (C2) holds trivially. Lemma A.1 above has shown that the identification condition in (C4) holds. Next, the continuity of $Q(b)$ is a result from Assumption SP4. To see this, first note that $Q(b)$ can be expressed as the sum of functions w.r.t. b of the following form: For some $(d_1,d_2)\in\{0,1\}^2$,

$$\begin{aligned}&P(y_{11}=d_1,y_{12}=d_2,x_{1(12)}^{(1)}+\tilde{x}_{1(12)}'\tilde{b}>0|\Omega)\\=&\int_{\tilde{x}_{1(12)}}\int_{-\tilde{x}_{1(12)}'\tilde{b}}P(y_{11}=d_1,y_{12}=d_2|x,\Omega)f_{x_{1(12)}^{(1)}|\tilde{x}_{1(12)},\Omega}(v)dvdF_{\tilde{x}_{1(12)}|\Omega}\end{aligned}$$

Then, $Q(b)$ is continuous if $f_{x_{1(12)}^{(1)}|\tilde{x}_{1(12)},\Omega}(\cdot)$ does not have any mass points, which is guaranteed by Assumption SP4.

The remaining task is to verify the uniform convergence condition (C3), i.e.,

$$\sup_{b\in\mathcal{B}}|Q_n(b)-Q(b)|=o_p(1)$$

This can be done by showing $\sup_{\mathcal{F}_n}|Q_n(b)-EQ_n(b)|=o_p(1)$ and $\sup_{b\in\mathcal{B}}|Q(b)-EQ_n(b)|=o(1)$, where $\mathcal{F}_n$ denotes the class of functions as $\mathcal{F}_n=\{K(z_1/h_n)\rho(b):b\in\mathcal{B}\}$.

First, note that $\mathcal{F}_n\subset\mathcal{F}=\{K(z_1/h)\rho(b):h>0,b\in\mathcal{B}\}=\mathcal{F}_h\times\mathcal{F}_b$ where $\mathcal{F}_h=\{K(z_1/h):h>0\}$ and $\mathcal{F}_b=\{\rho(b):b\in\mathcal{B}\}$. By Assumption SP8(i) and Lemma 22(ii) in Nolan and Pollard (1987), $\mathcal{F}_h$ is Euclidean for the constant envelope $\sup_{v\in\mathbb{R}^p}|K(v)|<$

∞. Then, as $\mathcal{F}_b$ is Euclidean for the constant envelope 1 (see Example 2.11 in Pakes and Pollard (1989)), $\mathcal{F}$ is Euclidean for the constant envelope $\sup_{v\in\mathbb{R}^p}|K(v)| < \infty$. Next, note that by Assumptions SP6 and SP8(ii),

$$\begin{aligned}\sup_{\mathcal{F}_n} E|K(z_1/h_n)\rho(b)| &= \sup_{\mathcal{F}_n}\int E[|K(z_1/h_n)\rho(b)||z_1]f(z_1)dz_1 \\ &= \sup_{\mathcal{F}_n} h_n^p\int K(v)E[|\rho(b)||z_1 = vh_n]f(vh_n)dv \\ &\leq \sup_{\mathcal{F}_n} h_n^p\int K(v)f(vh_n)dv = O(h_n^p)\end{aligned}$$

Then, under Assumption SP9(ii), applying Lemma 5 in Honoré and Kyriazidou (2000) yields

$$\sup_{\mathcal{F}_n} h_n^p|Q_n(b) - EQ_n(b)| = O_p\left(\sqrt{\frac{h_n^p\log n}{n}}\right) = o_p(h_n^p)$$

As the final step, we show that $\sup_{b\in\mathcal{B}}|Q(b) - EQ_n(b)| = o(1)$. Notice that by Assumptions SP7, SP8(ii), SP8(iii), and SP9(i),

$$\begin{aligned}\sup_{b\in\mathcal{B}}|Q(b) - EQ_n(b)| &= \sup_{b\in\mathcal{B}}|\varphi(0) - h_n^{-p}\int K(z_1/h_n)\varphi(z_1)dz_1| \\ &= \sup_{b\in\mathcal{B}}|\varphi(0) - h_n^{-p}\int K(z_1/h_n)[\varphi(0) + \varphi^{(1)}(\zeta)'z_1]dz_1| \\ &= \sup_{b\in\mathcal{B}}|\varphi(0) - \int K(v)[\varphi(0) + \varphi^{(1)}(v_n)'vh_n]dv| \\ &= \sup_{b\in\mathcal{B}}|h_n\int K(v)\varphi^{(1)}(v_n)'vdv| \\ &\leq h_n\sup_{b\in\mathcal{B}}\int K(v)|\varphi^{(1)}(v_n)|_1|v|_1dv \\ &= O(h_n) = o(1)\end{aligned}$$

where $|\cdot|_1$ denotes the l_1 norm of a vector. Therefore,

$$\sup_{b\in\mathcal{B}}|Q_n(b) - Q(b)| \leq \sup_{\mathcal{F}_n}|Q_n(b) - EQ_n(b)| + \sup_{b\in\mathcal{B}}|Q(b) - EQ_n(b)| = o_p(1)$$

which complete the proof. □

A.1.2 Rate of Convergence

Proof of Theorem 3.2. Denote $z = (z_1', z_2, z_3')'$. To facilitate exposition, we consider the following objective function of the estimator $\widehat{\beta}$:

$$g_{n,b}(z) = h_n^{-(J-1)p}K(z_1/h_n)z_2[sgn(z_3'b) - sgn(z_3'\beta_0)] = \kappa_n(z)(\mathbf{1}[z_3'b > 0] - \mathbf{1}[z_3'\beta_0 > 0])$$

where $\kappa_n(z) = 2h_n^{-(J-1)p}K(z_1/h_n)z_2$. By definition and change of variables, we have

$$E[\kappa_n(z)^2|z_3] = 4h_n^{-(J-1)p}\int K(v)^2 f_{z_1|z_2\neq 0,z_3}(vh_n)P(z_2\neq 0|z_3)dv$$

almost surely for all n. Under Assumptions SP3, SP6', and SP8'(i), there exist some $c_1, c_2 > 0$ such that $c_1 < h_n^{(J-1)p}E[\kappa_n(z)^2|z_3] < c_2$ almost surely. Then, using the same argument in Seo and Otsu (2018) (Section B.1 of the supplementary material), we have for all $b_1, b_2 \in \mathcal{B}$,

$$\begin{aligned}
&h_n^{(J-1)p/2}\|(g_{n,b_1}(z) - g_{n,b_2}(z)\|_2 \\
=&E[h_n^{(J-1)p}\kappa_n(z)^2(\mathbf{1}[z_3'b_1 > 0] - \mathbf{1}[z_3'b_2 > 0])^2]^{1/2} \\
=&E[h_n^{(J-1)p}E[\kappa_n(z)^2|z_3](\mathbf{1}[z_3'b_1 > 0] - \mathbf{1}[z_3'b_2 > 0])^2]^{1/2} \\
=&\geq c_1^{1/2}E|\mathbf{1}[z_3'b_1 > 0] - \mathbf{1}[z_3'b_2 > 0]| \asymp |b_1 - b_2|_2
\end{aligned} \tag{A.1}$$

where $\|\cdot\|_2$ denotes the $L_2(P)$ norm. Similarly, we can obtain

$$\begin{aligned}
&h_n^{(J-1)p}E[\sup_{b\in\mathcal{B}:|b-\beta|_2<\varepsilon}|g_{n,b}(z) - g_{n,\beta}(z)|^2] \\
=&E[h_n^{(J-1)p}E[|\kappa_n(z)|^2|z_3]\sup_{b\in\mathcal{B}:|b-\beta|_2<\varepsilon}|\mathbf{1}[z_3'b > 0] - \mathbf{1}[z_3'\beta > 0)|^2] \\
\leq&c_2E\sup_{b\in\mathcal{B}:|b-\beta|_2<\varepsilon}|\mathbf{1}[z_3'b > 0] - \mathbf{1}[z_3'\beta > 0]| \leq c_2'\varepsilon
\end{aligned} \tag{A.2}$$

for some $c_2' > 0$, sufficiently large n, and all β in a neighborhood of β_0.

Next, note that under Assumptions SP8'(ii)-(iv), SP7', and SP9'(iii), we have

$$\begin{aligned}
E[g_{n,b}(z)] =& \int K(v)E[z_2(sgn(z_3'b) - sgn(z_3'\beta_0))|z_1 = vh_n]f_{z_1}(vh_n)dv \\
=&f_{z_1}(0)E[z_2(sgn(z_3'b) - sgn(z_3'\beta_0))|z_1 = 0] \\
&+ h_n^2\int K(v)v'\frac{\partial^2 f_{z_1}(\tau)E[z_2(sgn(z_3'b) - sgn(z_3'\beta_0))|z_1 = \tau]}{\partial\tau\partial\tau'}|_{\tau=\bar{v}}vdv \\
=&f_{z_1}(0)E[z_2(sgn(z_3'b) - sgn(z_3'\beta_0))|z_1 = 0] + o((nh_n^{(J-1)p})^{2/3})
\end{aligned} \tag{A.3}$$

where $\bar{v}$ is a point on the line joining 0 and vh_n, and the second equality follows from the dominated convergence theorem and mean value theorem.

Denote $\mathcal{Z}_b = \{z_3 : sgn(z_3'b) \neq sgn(z_3'\beta_0)\}$ for all $b \in \mathcal{B}\setminus\{\beta_0\}$. Following similar argument used in the proof of Lemma A.4,

$$\begin{aligned}
-E[z_2(sgn(z_3'b) - sgn(z_3'\beta_0))|z_1 = 0] &= 2\int_{\mathcal{Z}_b} sgn(z_3'\beta_0)E[z_2|z_3, z_1 = 0]dF_{z_3|z_1=0} \\
&= 2\int_{\mathcal{Z}_b}|E[z_2|z_3, z_1 = 0]|dF_{z_3|z_1=0} > 0
\end{aligned}$$

Therefore, applying the same argument as Kim and Pollard (1990) pp. 214-215 yields

$$\frac{\partial}{\partial b}E[z_2(sgn(z_3'b)|z_1=0]|_{b=\beta_0}=0 \tag{A.4}$$

and

$$\begin{aligned}&-\frac{\partial^2 E[z_2(sgn(z_3'b)-sgn(z_3'\beta_0))|z_1=0]}{\partial b\partial b'}\\&=\int \mathbf{1}[z_3'\beta_0=0]\left(\frac{\partial}{\partial z_3}E[z_2|z_3,z_1=0]\right)'\beta_0 z_3 z_3' f_{z_3|z_1=0}(z_3)d\mu_{\beta_0}\end{aligned} \tag{A.5}$$

where μ_{β_0} is the surface measure on the boundary of $\{z_3 : z_3'\beta_0 \geq 0\}$.

Putting (A.3), (A.4), and (A.5) together, we have

$$E[g_{n,\beta}(z)]=\frac{1}{2}(b-\beta_0)'V(b-\beta_0)+o(|b-\beta_0|_2^2)+o((nh_n^{(J-1)p})^{2/3}) \tag{A.6}$$

where

$$V=-2f_{z_1}(0)\int \mathbf{1}[z_3'\beta_0=0]\left(\frac{\partial}{\partial z_3}E[z_2|z_3,z_1=0]\right)'\beta_0 z_3 z_3' f_{z_3|z_1=0}(z_3)d\mu_{\beta_0}$$

Notice that $h_n^{(J-1)p}g_{n,b}(z)$ is uniformly bounded by Assumption SP8'(i) and $\lim_{n\to\infty}Eg_{n,b}(z)$ is uniquely maximized at β_0 by Lemma A.1. Then, putting (A.1), (A.2), and (A.6) together, by Lemma 1 of Seo and Otsu (2018), we can conclude that there exists some positive constant C for each $\varepsilon>0$ such that

$$\left|\frac{1}{n}\sum_{i=1}^n g_{n,b}(z_i)-E[g_{n,b}(z)]\right|\leq \varepsilon|b-\beta_0|_2^2+O_p\left(\left(nh_n^{(J-1)p}\right)^{-2/3}\right) \tag{A.7}$$

for all $b\in\{\mathcal{B}:(nh_n^{(J-1)p})^{-1/3}\leq|b-b_0|_2\leq C\}$. Then, assuming $|b-b_0|_2\geq(nh_n^{(J-1)p})^{-1/3}$, we have, by (A.7) and (A.3),

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n g_{n,\widehat{\beta}}(z_i)&\leq E[g_{n,\widehat{\beta}}(z)]+\varepsilon|\widehat{\beta}-\beta_0|_2^2+O_p((nh_n^{(J-1)p})^{-2/3})\\&\leq(\varepsilon-C')|\widehat{\beta}-\beta_0|_2^2+o(|\widehat{\beta}-\beta_0|_2^2)+O_p((nh_n^{(J-1)p})^{-2/3})\end{aligned} \tag{A.8}$$

for each $\varepsilon>0$ and some positive constant C'. By the definitions of $\widehat{\beta}$ and $g_{n,b}(\cdot)$,

$$\frac{1}{n}\sum_{i=1}^n g_{n,\widehat{\beta}}(z_i)=\frac{1}{n}\sum_{i=1}^n g_{n,\widehat{\beta}}(z_i)-\frac{1}{n}\sum_{i=1}^n g_{n,\beta_0}(z_i)\geq o_p((nh_n^{(J-1)p})^{-2/3}) \tag{A.9}$$

Then, the desired result follows from taking ε sufficiently small such that $\varepsilon-C'<0$ and combining (A.8) and (A.9). □

A.2 Dynamic Panel Data Estimators

Here, we only establish regularity conditions and prove consistency of our dynamic panel data estimator, as consistency for the static model follows as a special case.

Consider the events:

$$A = \{y_{10} = d_0, y_{11} = 1, y_{12} = 0, y_{13} = d_3\}$$
$$B = \{y_{10} = d_0, y_{11} = 0, y_{12} = 1, y_{13} = d_3\}$$

where d_0 and d_3 are either 0 or 1. In what follows, denote $z = (x'_{1(12)}, y_{1(03)})'$.

Lemma A.2. *Under Assumption DP3,* $sgn(P(A|x,\alpha,\Omega) - P(B|x,\alpha,\Omega)) = sgn(z'\theta_0)$.

Proof of Lemma B.1. By Assmption DP3, we have

$$\begin{aligned}
P(A|x,\alpha,\Omega) =& P(y_{10} = 1|x,\alpha,\Omega)^{d_0}(1 - P(y_{10} = 1|x,\alpha,\Omega))^{1-d_0} \\
&\times P(x'_{11}\beta_0 + \gamma_0 d_0 + \alpha_1 - \epsilon_{11} > \max\{x'_{21}\beta_0 + \alpha_2 - \epsilon_{21}, 0\}|x,\alpha,\Omega) \\
&\times (1 - P(x'_{12}\beta_0 + \gamma_0 + \alpha_1 - \epsilon_{12} > \max\{x'_{21}\beta_0 + \alpha_2 - \epsilon_{22}, 0\}|x,\alpha,\Omega)) \\
&\times P(x'_{12}\beta_0 + \alpha_1 - \epsilon_{13} > \max\{x'_{21}\beta_0 + \alpha_2 - \epsilon_{23}, 0\}|x,\alpha,\Omega)^{d_3} \\
&\times (1 - P(x'_{12}\beta_0 + \alpha_1 - \epsilon_{13} > \max\{x'_{21}\beta_0 + \alpha_2 - \epsilon_{23}, 0\}|x,\alpha,\Omega))^{1-d_3}
\end{aligned}$$

and similarly,

$$\begin{aligned}
P(B|x,\alpha,\Omega) =& P(y_{10} = 1|x,\alpha,\Omega)^{d_0}(1 - P(y_{10} = 1|x,\alpha,\Omega))^{1-d_0} \\
&\times (1 - P(x'_{11}\beta_0 + \gamma_0 d_0 + \alpha_1 - \epsilon_{11} > \max\{x'_{21}\beta_0 + \alpha_2 - \epsilon_{21}, 0\}|x,\alpha,\Omega)) \\
&\times P(x'_{12}\beta_0 + \alpha_1 - \epsilon_{12} > \max\{x'_{21}\beta_0 + \alpha_2 - \epsilon_{22}, 0\}|x,\alpha,\Omega) \\
&\times P(x'_{12}\beta_0 + \gamma_0 + \alpha_1 - \epsilon_{13} > \max\{x'_{21}\beta_0 + \alpha_2 - \epsilon_{23}, 0\}|x,\alpha,\Omega)^{d_3} \\
&\times (1 - P(x'_{12}\beta_0 + \gamma_0 + \alpha_1 - \epsilon_{13} > \max\{x'_{21}\beta_0 + \alpha_2 - \epsilon_{23}, 0\}|x,\alpha,\Omega))^{1-d_3}
\end{aligned}$$

It is not hard to verify that

$$\frac{P(A|x,\alpha,\Omega)}{P(B|x,\alpha,\Omega)} > 1 \Leftrightarrow x'_{11}\beta_0 + \gamma_0 d_0 > x'_{12}\beta_0 + \gamma_0 d_3$$

for each of the 4 cases corresponding to the values of d_0 and d_3. Then, the desired result follows. □

In what follows, we focus on the case where $\Theta \subset \{\theta = (b', g)' \in \mathbb{R}^{p+1} : b_1 = 1\}$. The case with $\Theta \subset \{\theta = (b', g)' \in \mathbb{R}^{p+1} : b_1 = -1\}$ is symmetric.

Lemma A.3. *Under Assumptions DP3 - DP5,* $P(sgn(z'\theta) \neq sgn(z'\theta_0)|\Omega) > 0$ *for all* $\theta \in \Theta \setminus \{\theta_0\}$.

Proof of Lemma B.2. To prove the statement in the lemma, it suffices to show that for all $\theta \in \Theta \setminus \{\theta_0\}$, (i) $P(\tilde{z}'\tilde{\theta} \neq \tilde{z}'\tilde{\theta}_0|\Omega) > 0$, and (ii) $P(x^{(1)}_{1(12)} \in \mathcal{N}|\tilde{x}_{1(12)}, y_{10} = d_0, y_{13} = d_3, \Omega) > 0$ for all $(d_0, d_3) \in \{0,1\}^2$ and for any proper interval $\mathcal{N}$ on the real line.

(i) If $g = \gamma_0$, then $P(\tilde{z}'\tilde{\theta} = \tilde{z}'\tilde{\theta}_0|\Omega) = P(\tilde{x}'_{1(12)}(\tilde{b} - \tilde{\beta}_0) = 0|\Omega) < 1$ by DP5. For the case with $g \neq \gamma_0$,

$$\begin{aligned}
&P(\tilde{z}'\tilde{\theta} = \tilde{z}'\tilde{\theta}_0|\Omega) \\
&= \sum_{d_0 \in \{0,1\}} \int P((g-\gamma_0)y_{13} = (\gamma_0 - g)d_0 + \tilde{x}'_{1(12)}(\tilde{\beta}_0 - \tilde{b})|y_{10} = d_0, \tilde{x}_{1(12)}, \Omega) \\
&\quad \times P(y_{10} = d_0|\tilde{x}_{1(12)}, \Omega) dF_{\tilde{x}_{1(12)}|\Omega}
\end{aligned}$$

By Assumption DP3, $P((g-\gamma_0)y_{13} = (\gamma_0 - g)d_0 + \tilde{x}'_{1(12)}(\tilde{\beta}_0 - \tilde{b})|y_{10} = d_0, \tilde{x}_{1(12)}, \Omega) < 1$ for all $d_0 \in \{0,1\}$, and hence $P(\tilde{z}'\tilde{\theta} \neq \tilde{z}'\tilde{\theta}_0|\Omega) > 0$.

(ii) For any given d_0 and d_3, by Bayes' theorem,

$$\begin{aligned}
&P(x^{(1)}_{1(12)} \in \mathcal{N}|\tilde{x}_{1(12)}, y_{10} = d_0, y_{13} = d_3, \Omega) \\
=&\frac{P(y_{10} = d_0, y_{13} = d_3|\tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) P(x^{(1)}_{1(12)} \in \mathcal{N}|\tilde{x}_{1(12)}, \Omega)}{P(y_{10} = d_0, y_{13} = d_3|\tilde{x}_{1(12)}, \Omega)}
\end{aligned}$$

Assumption DP4 guarantees that $P(x^{(1)}_{1(12)} \in \mathcal{N}|\tilde{x}_{1(12)}, \Omega) > 0$. Furthermore, note that

$$\begin{aligned}
&P(y_{10} = d_0, y_{13} = d_3|\tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) \\
=&\int P(y_{13} = d_3|x, \alpha, y_{10} = d_0, \tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) \\
&\quad \times P(y_{10} = d_0|x, \alpha, \tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) dF_{x,\alpha|\tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega} \\
=&\sum_{(d_1,d_2) \in \{0,1\}^2} \int P(y_{13} = d_3|x, \alpha, y_{10} = d_0, y_{11} = d_1, y_{12} = d_2, \tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) \\
&\quad \times P(y_{12} = d_2|x, \alpha, y_{10} = d_0, y_{11} = d_1, \tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) \\
&\quad \times P(y_{11} = d_1|x, \alpha, y_{10} = d_0, \tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) \\
&\quad \times P(y_{10} = d_0|x, \alpha, \tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) dF_{x,\alpha|\tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega}
\end{aligned}$$

Therefore, $P(y_{10} = d_0, y_{13} = d_3|\tilde{x}_{1(12)}, x^{(1)}_{1(12)} \in \mathcal{N}, \Omega) > 0$ by Assumption DP3, and hence $P(x^{(1)}_{1(12)} \in \mathcal{N}|\tilde{x}_{1(12)}, y_{10} = d_0, y_{13} = d_3, \Omega) > 0$.

Then, the desired result follows by letting $\mathcal{N} = \{-\tilde{z}'\tilde{\theta}_0 < x^{(1)}_{1(12)} < -\tilde{z}'\tilde{\theta}\} \cup \{-\tilde{z}'\tilde{\theta} < x^{(1)}_{1(12)} < -\tilde{z}'\tilde{\theta}_0\}$. □

Define the population objective function:

$$Q(\theta) = f(0)E[\psi(\theta)|\Omega]$$

The following lemma establish the point identification of θ_0 relative to all $\theta \in \Theta \setminus \{\theta_0\}$.

Lemma A.4. *Under Assumptions DP3 - DP6, $Q(\theta_0) > Q(\theta)$ for all $\theta \in \Theta \setminus \{\theta_0\}$.*

Proof of Lemma B.3. Recall that $\psi(\theta) = y_{1(12)}sgn(x'_{1(12)}b + gy_{1(03)}) = y_{1(12)}sgn(z'\theta)$ for all $\theta \in \Theta$. Let $\mathcal{Z}_\theta \equiv \{z : sgn(z'\theta) \neq sgn(z'\theta_0), \theta \in \Theta \setminus \{\theta_0\}\}$. Lemma A.3 shows that $P(\mathcal{Z}_\theta|\Omega) > 0$. Then, by definition,

$$\begin{aligned}
&Q(\theta_0) - Q(\theta) \\
=&f(0)E[y_{1(12)}(sgn(z'\theta_0) - sgn(z'\theta))|\Omega] \qquad \text{(A.10)} \\
=&2f(0)\int_{\mathcal{Z}_\theta} sgn(z'\theta_0)E[y_{1(12)}|z,\Omega]dF_{z|\Omega} \\
=&2f(0)\int_{\mathcal{Z}_\theta} sgn(z'\theta_0)E[E[y_{1(12)}|x,\alpha,y_{10}=d_0,y_{13}=d_3,\Omega]|z,\Omega]dF_{z|\Omega} \\
=&2f(0)\int_{\mathcal{Z}_\theta} sgn(z'\theta_0)E[E[\mathbf{1}[y_{11}=1,y_{12}=0]|x,\alpha,y_{10}=d_0,y_{13}=d_3,\Omega] \\
&- E[\mathbf{1}[y_{11}=0,y_{12}=1]|x,\alpha,y_{10}=d_0,y_{13}=d_3,\Omega]|z,\Omega]dF_{z|\Omega} \\
=&2f(0)\int_{\mathcal{Z}_\theta} sgn(z'\theta_0)E[P(y_{11}=1,y_{12}=0|x,\alpha,y_{10}=d_0,y_{13}=d_3,\Omega) \\
&- P(y_{11}=0,y_{12}=1|x,\alpha,y_{10}=d_0,y_{13}=d_3,\Omega)|z,\Omega]dF_{z|\Omega} \\
=&2f(0)\int_{\mathcal{Z}_\theta} E\left[sgn(z'\theta_0)\left(\frac{P(A|x,\alpha,\Omega)-P(B|x,\alpha,\Omega)}{P(y_{10}=d_0,y_{13}=d_3|x,\alpha,\Omega)}\right)|z,\Omega\right]dF_{z|\Omega}
\end{aligned}$$

It follows from Lemma A.2 that

$$sgn(z'\theta_0)\left(\frac{P(A|x,\alpha,\Omega)-P(B|x,\alpha,\Omega)}{P(y_{10}=d_0,y_{13}=d_3|x,\alpha,\Omega)}\right) \geq 0$$

and hence

$$\begin{aligned}
&E\left[sgn(z'\theta_0)\left(\frac{P(A|x,\alpha,\Omega)-P(B|x,\alpha,\Omega)}{P(y_{10}=d_0,y_{13}=d_3|x,\alpha,\Omega)}\right)|z,\Omega\right] \\
=&E\left[\left|sgn(z'\theta_0)\left(\frac{P(A|x,\alpha,\Omega)-P(B|x,\alpha,\Omega)}{P(y_{10}=d_0,y_{13}=d_3|x,\alpha,\Omega)}\right)\right||z,\Omega\right] \\
=&E\left[\left|\frac{P(A|x,\alpha,\Omega)-P(B|x,\alpha,\Omega)}{P(y_{10}=d_0,y_{13}=d_3|x,\alpha,\Omega)}\right||z,\Omega\right] \geq 0
\end{aligned}$$

Note that the expectation above is strictly positive for almost all z since $P(A|x,\alpha,\Omega) - P(B|x,\alpha,\Omega) = 0$ if and only if $sgn(z'\theta_0) = 0$ which is an event having zero probability measure under Assumption DP4. It then follows from Lemma A.3 and Assumption DP6 that $Q(\theta_0) - Q(\theta) > 0$ for all $\theta \in \Theta \setminus \{\theta_0\}$. □

To simplify notation, we define

$$Q_n(\theta) = \frac{1}{nh_n^{3p}} \sum_{i=1}^{n} K(\Delta x_i / h_n)\psi_i(\theta)$$

and

$$\phi(\cdot) = f(\cdot)E[\psi(\theta)|x_{2(12)} = x_{2(23)} = x_{1(23)} = \cdot]$$

Proof of Theorem 3.3. The proof proceeds by verifying the following conditions for Theorem 9.6.1 in Amemiya (1985): (C1) Θ is a compact set, (C2) $Q_n(\theta)$ is a measurable function for all $\theta \in \Theta$, (C3) $Q_n(\theta)$ converges in probability to a nonstochastic function $Q(\theta)$ uniformly in $\theta \in \Theta$, (C4) $Q(\theta)$ is continuous in θ and is uniquely maximized at θ_0.

The compactness of Θ is satisfied by construction. Condition (C2) holds trivially. Lemma B.3 above has shown that the identification condition in (C4) holds. Next, the continuity of $Q(\theta)$ is guaranteed by Assumptions DP3 and DP4. To see this, first note that $Q(\theta)$ can be expressed as the sum of functions (with respect to θ) of the following form:

$$\begin{aligned}
&P(y_{11} = d_1, y_{12} = d_2, x^{(1)}_{1(12)} + \tilde{x}'_{1(12)}\tilde{b} + g(y_{10} - y_{13}) > 0|\Omega) \\
&= \sum_{(d_0,d_3)\in\{0,1\}^2} \int_{\tilde{x}_{1(12)}} \left[\int_{-\tilde{x}'_{1(12)}\tilde{b} - g(d_0 - d_3)} P(y_{11} = d_1, y_{12} = d_2|x, y_{10} = d_0, y_{13} = d_3, \Omega) \right. \\
&\left. \times f_{x^{(1)}_{1(12)}|\tilde{x}_{1(12)}, y_{10}=d_0, y_{13}=d_3, \Omega}(v)dv \right] dF_{\tilde{x}_{1(12)}|y_{10}=d_0, y_{13}=d_3, \Omega} \times P(y_{10} = d_0, y_{13} = d_3|\Omega)
\end{aligned}$$

Then, the function above is continuous if $f_{x^{(1)}_{1(12)}|\tilde{x}_{1(12)}, y_{10}=d_0, y_{13}=d_3, \Omega}(\cdot)$ does not have any mass points, which is implied by Assumptions DP3, DP4, and Bayes' theorem.

The remaining step is to verify the uniform convergence condition (C3), i.e.,

$$\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| = o_p(1)$$

This can be done by showing $\sup_{\mathcal{F}_n} |Q_n(\theta) - EQ_n(\theta)| = o_p(1)$ and $\sup_{\theta\in\Theta} |EQ_n(\theta) - Q(\theta)| = o(1)$, where $\mathcal{F}_n$ denotes the class of functions as $\mathcal{F}_n = \{K(\Delta x/h_n)\psi(\theta) : \theta \in \Theta\}$.

It is clear that $\mathcal{F}_n \subset \mathcal{F} \equiv \{K(\Delta x/h)\psi(\theta) : h > 0, \theta \in \Theta\} = \mathcal{F}_h \times \mathcal{F}_\theta$ with $\mathcal{F}_h \equiv \{K(\Delta x/h) : h > 0\}$ and $\mathcal{F}_\theta \equiv \{\psi(\theta) : \theta \in \Theta\}$. By Assumption DP8(i) and Lemma 22(ii) in

Nolan and Pollard (1987), $\mathcal{F}_h$ is Euclidean for the constant envelope $\sup_{v\in\mathbb{R}^{3p}}|K(v)| < \infty$. Furthermore, as $\mathcal{F}_\theta$ is Euclidean for the constant envelope $\sup_{\theta\in\Theta}|\psi(\theta)| = 1$ (see Example 2.11 in Pakes and Pollard (1989)), $\mathcal{F}$ is Euclidean for the constant envelope $\sup_{v\in\mathbb{R}^{3k}}|K(v)| < \infty$. Next, note that by Assumptions DP6 and DP8(ii),

$$\begin{aligned}\sup_{\mathcal{F}_n} E|K(\Delta x/h_n)\psi(\theta)| &= \sup_{\mathcal{F}_n}\int [E|K(\Delta x/h_n)\psi(\theta)||\Delta x]f(\Delta x)d\Delta x \\ &= \sup_{\mathcal{F}_n} h_n^{3p}\int K(v)[E|\psi(\theta)||\Delta x = vh_n]f(vh_n)dv \\ &\leq \sup_{\mathcal{F}_n} h_n^{3p}\int K(v)f(vh_n)dv = O(h_n^{3p})\end{aligned}$$

Then, under Assumption DP9(ii), we obtain by applying Lemma 5 in Honoré and Kyriazidou (2000) that

$$\sup_{\mathcal{F}_n} h_n^{3p}|Q_n(\theta) - EQ_n(\theta)| = O_p\left(\sqrt{\frac{h_n^{3p}\log n}{n}}\right) = o_p(h_n^{3p})$$

Next, we show that $\sup_{\theta\in\Theta}|EQ_n(\theta) - Q(\theta)| = o(1)$. Notice that by Assumptions DP7, DP8(ii), DP8(iii), and DP9(i),

$$\begin{aligned}\sup_{\theta\in\Theta}|EQ_n(\theta) - Q(\theta)| &= \sup_{\theta\in\Theta}|\frac{1}{h_n^{3p}}\int K(\Delta x/h_n)\phi(\Delta x)d\Delta x - \phi(0)| \\ &= \sup_{\theta\in\Theta}|\frac{1}{h_n^{3p}}\int K(\Delta x/h_n)[\phi(0) + \phi^{(1)}(\zeta)'\Delta x]d\Delta x - \phi(0)| \\ &= \sup_{\theta\in\Theta}|\phi(0)\int K(v)dv + h_n\int K(v)\phi^{(1)}(v_n)'vdv - \phi(0)| \\ &= \sup_{\theta\in\Theta}|h_n\int K(v)\phi^{(1)}(v_n)'vdv| \\ &\leq h_n\int K(v)|\phi^{(1)}(v_n)|_1|v|_1dv \\ &= O(h_n) = o(1)\end{aligned}$$

where $|\cdot|_1$ denotes the l_1 norm of a vector. Therefore,

$$\sup_{\theta\in\Theta}|Q_n(\theta) - Q(\theta)| \leq \sup_{\mathcal{F}_n}|Q_n(\theta) - EQ_n(\theta)| + \sup_{\theta\in\Theta}|EQ_n(\theta) - Q(\theta)| = o_p(1)$$

and the desired result follows. □