Inference on Semiparametric Multinomial Response Models*

Shakeeb Khan† Boston College

Fu Ouyang‡ University of Queensland

Elie Tamer§ Harvard University

December 4, 2020

Abstract

We explore inference on regression coefficients in semiparametric multinomial response models. We consider cross-sectional, and both static and dynamic panel settings where we focus throughout on inference under sufficient conditions for point identification. The approach to identification uses a matching insight throughout all three models coupled with 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 generalize the work of Honoré and Kyriazidou (2000) to cover the multinomial case. A simulation study establishes adequate finite sample properties of our new procedures. We apply our estimators to a scanner panel data set.

Keywords: Multinomial Response, Rank Estimation, Dynamic Panel Data.

1 Introduction

Many important economic decisions involve households' or firms' choice among qualitative or discrete alternatives. Examples are individuals' choice among transportation alternatives, family

*We thank Co-Editor Andres Santos and two anonymous referees for comments that improved the content and exposition of the paper. We are also grateful for helpful comments received from conference participants at the 2016 Panel Data Workshop at University of Amsterdam, the 22nd Panel Data Conference in Perth, Australia, the 2016 Australasia Meeting of the Econometric Society at UTS, and seminar participants at various institutions. The usual disclaimer applies.

†Email: shakeeb.khan@bc.edu.

‡Email: f.ouyang@uq.edu.au.

§Email: elietamer@fas.harvard.edu.

sizes, residential locations, brands of automobiles, health plans, etc. The theory of discrete choice is designed to model these kinds of choice settings and to provide the corresponding econometric methodology that allows for model estimation and prediction. A standard approach in the econometrics discrete choice literature is to model discrete choices 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 individuals' preferences and the choice constraints. As in most of the econometrics literature, individuals 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 is interested in learning these preferences for the purpose of prediction and counterfactual analysis. Given a parametric model for preferences, the objective of this paper is to learn the finite-dimensional coefficients that characterize this model using multinomial choice data.

There has been a renewed interest recently among applied economists in estimating models of multinomial choice with both cross-section and panel data. In marketing, IO, and other literature, 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 models. For example, Ahn, Powell, Ichimura, and Ruud (2017) study 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-sectional and panel data multinomial response models where we use a unified approach for identification in all three classes of multinomial models: cross-sectional, static panel, and dynamic panel. Throughout and most importantly, we relax the *IIA property* by allowing for arbitrary correlation in the unobserved errors across choices. This IIA property, that is often used in applications, can result in unintuitive subsititution patterns. 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 models and require different variations in the regressors over time to garner point identification. The contribution here is a model that extends the binary choice panel data

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

model in Manski (1987) to multinomial response models. We derive large sample properties of the multinomial maximum score (MS) estimator and show that its rate of convergence 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 asymptotics. This generalizes the work of Honoré and Kyriazidou (2000) to multinomial settings and complements their work by providing rate and large-sample distribution results. Our approaches for both cross-sectional and panel data models are robust in the sense that they achieve meaningful bounds for the preference 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 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 alternative specific effects. For this model, we propose a localized maximum score (analogous to Manski (1987)) estimator and show its point consistency, rate of convergence, and limiting distribution under mild regularity conditions. 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, Ponomareva, and Tamer (2019). Here again, we establish large sample properties of our procedure 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. A supplementary appendix collects all proofs.

For ease of reference, the notation adopted in the next sections of the paper are listed here:

Notation. In the sections that follow, where variables will depend on the individual, choice and time period, we will use letters i and m for indexing individuals, j and k for indexing alternatives, and s and t for indexing time periods. Further notation we will adopt to help clarification include having the first element of a (constant or random) vector ν be denoted by $\nu^{(1)}$ and the sub-vector comprising its remaining elements be denoted by $\tilde{\nu}$. $\mathbf{1}[\cdot]$ denotes the indicator function that equals 1 when the event in the brackets occurs, and 0 otherwise. For two random vectors u and v, the notation $u \overset{d}{=} v|\cdot$ means that u and v have identical distribution conditional on $\cdot$, and $u \perp v|\cdot$

means that u and v are independent conditional on $\cdot$. We use $F_u(\cdot)$ ($F_{u|v}(\cdot)$) and $f_u(\cdot)$ ($f_{u|v}(\cdot)$) to denote the joint cumulative distribution function (CDF) and probability density function (PDF) of u (conditional on v), respectively. The interior of a set S is denoted by $\text{int}(S)$ and the symbol $\setminus$ represents set difference.

2 Cross-Sectional Multinomial Choice

2.1 Semiparametric Multinomial Choice

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

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

with the convention that $y_{ij} = 0$ indicates that the choice of alternative j is not made by individual i. As is standard in the literature, an assumption of joint continuity of the indirect utilities rules out ties (with probability one). In addition, we maintain the familiar linear form for indirect utilities[2]

$$\begin{aligned} y_{i0}^* &= 0, \\ y_{ij}^* &= x_{ij}'\beta_0 - \epsilon_{ij}, \ j = 1, ..., J, \end{aligned} \tag{2.1}$$

where β_0 is a p-dimensional vector of unknown preference 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 $\epsilon_i \equiv (\epsilon_{i1}, ..., \epsilon_{iJ})'$ of unobserved error terms, attained by stacking all the scalar idiosyncratic errors ϵ_{ij}, is assumed to be jointly continuously distributed and independent of the $p \times J$-dimensional vector of regressors $x_i \equiv (x_{i1}', ..., x_{iJ}')'$.[3] We stress that expression (2.1) is rather general. By properly re-organizing x_{ij}'s and β_0, (2.1) can accommodate both alternative-specific and individual-specific covariates.[4]

Parametric assumptions on the unobservables, such as i.i.d. Type I extreme value (multinomial

[2] Our method can be applied to more general models with indirect utilities $y_{ij}^* = u_j(x_{ij}'\beta_0, -\epsilon_{ij})$, $j = 1, 2$, where $u_j(\cdot, \cdot)$'s are unknown (to econometrician) $\mathbb{R}^2 \mapsto \mathbb{R}$ functions strictly increasing in each of their arguments. It will be clear that our rank procedure does not rely on the additive separability of the regressors and error terms.

[3] We impose the independence restriction here to simplify exposition. As will become clear below, our matching-based approach allows ϵ_i to be correlated with individual-specific regressors.

[4] See Cameron and Trivedi (2005) p. 498 for a detailed discussion. But the identification of models with both alternative-specific and individual-specific regressors will need to take two steps, of which the first step only identifies the coefficients on alternative-specific regressors. See Remarks 1 and 2 below.

Logit) or multivariate normal (multinomial Probit), have been used to attain identification.[5] 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 have been approaches to ameliorate these problems by, for example, using nested Logit models, and simulation-based approaches have been successfully used to approximate multiple integrals.

We take another approach. This paper is 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. Previous contributions to this question include Lee (1995), who proposes a profile likelihood approach, extending the results in Klein and Spady (1993) for the binary response model. Ahn, Powell, Ichimura, and Ruud (2017) propose a two-step estimator that requires nonparametric methods but show the second step is of closed-form. Shi, Shum, and Song (2018) also propose a two-step estimator in panel setups exploiting a cyclic monotonicity condition, which also requires a high dimensional nonparametric first stage, but whose second stage is not closed-form as Ahn, Powell, Ichimura, and Ruud (2017) is.

The next section demonstrates the main intuition that runs through the various models in this paper. It is provided for the cross-sectional multinomial response model.

2.2 Local Rank Procedure

Consider a multinomial response model with 3 alternatives ($J = 2$) for now where the indirect utilities for alternatives 0, 1, and 2 are

$$\begin{aligned} y_{i0}^* &= 0, \\ y_{ij}^* &= x_{ij}'\beta_0 - \epsilon_{ij}, \ j = 1, 2. \end{aligned}$$

This simple model is sufficient to illustrate our approach, which is straightforward to be applied to data with more alternatives.

Given the indirect utilities, the observed dependent variables y_{ij} is of the form

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

[5] In these parametric models, the indirect utilities are typically specified as $u_{ij}^* = w_{ij}'\beta_0 - e_{ij}$, $j = 0, 1, ..., J$ and the corresponding distributional restrictions are imposed on errors $e_i \equiv (e_{i0}, ..., e_{iJ})'$. Expression (2.1) can be written from these models by location normalization $x_{ij} \equiv w_{ij} - w_{i0}$ and $\epsilon_{ij} \equiv e_{ij} - e_{i0}$.

and the choice probabilities, for a given β_0 and joint distribution F_ϵ on the ϵ_i, are expressed by

$$G(x_i;\beta_0,F_\epsilon)=\left(\begin{array}{c} P(x_{i1}'\beta_0-\epsilon_{i1}<0,x_{i2}'\beta_0-\epsilon_{i2}<0) \\ P(x_{i1}'\beta_0-\epsilon_{i1}>0,x_{i1}'\beta_0-\epsilon_{i1}>x_{i2}'\beta_0-\epsilon_{i2}) \\ P(x_{i1}'\beta_0-\epsilon_{i1}<x_{i2}'\beta_0-\epsilon_{i2},x_{i2}'\beta_0-\epsilon_{i2}>0) \end{array}\right). \tag{2.2}$$

Assuming a random sample of $\{y_{i0},y_{i1},y_{i2},x_i\}_{i=1}^n$, we are interested in the identification of β_0. In what follows, we propose a procedure attaining point identification of β_0 when there is at least one continuous regressor with large support, and attaining bounds for β_0 when all regressors are discrete. To this end, we maintain the assumption that $\epsilon_i \perp x_i$ but allow for arbitrary correlation between ϵ_{i1} and ϵ_{i2}.

To illustrate how we garner information about β_0 from model (2.2), 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*.[6] By this we mean that conditional on x_i, $P(y_{i1}=1|x_{i1},x_{i2}=x_2)$ is increasing in $x_{i1}'\beta_0$ for all constant vector x_2. Thus for all $i\neq m$, we have the following (conditional) identification inequality

$$P(y_{i1}=1|x_i,x_{i2}=x_2)\geq P(y_{m1}=1|x_m,x_{m2}=x_2)\Leftrightarrow x_{i1}'\beta_0\geq x_{m1}'\beta_0. \tag{2.3}$$

Fixing regressors of all other alternatives to obtain monotone index models for one alternative motivates all our identification results in this paper.

Note that the monotonic relation specified in (2.3) can be repeated for all values of x_2 (finitely many if the support of x_{i2} is finite). Besides, note that for a fixed x_2, $P(y_{i0}=1|x_i,x_{i2}=x_2)$ and $P(y_{i2}=1|x_i,x_{i2}=x_2)$ are both decreasing in $x_{i1}'\beta_0$, which gives additional identification inequalities for alternative 1, i.e.,

$$P(y_{i0}=1|x_i,x_{i2}=x_2)\leq P(y_{m0}=1|x_m,x_{m2}=x_2)\Leftrightarrow x_{i1}'\beta_0\geq x_{m1}'\beta_0$$

and

$$P(y_{i2}=1|x_i,x_{i2}=x_2)\leq P(y_{m2}=1|x_m,x_{m2}=x_2)\Leftrightarrow x_{i1}'\beta_0\geq x_{m1}'\beta_0.$$

Furthermore, similar conditional monotone index model can also be exploited by fixing x_{i1} at some constant vector x_1, resulting the following identification inequalities

$$P(y_{i2}=1|x_i,x_{i1}=x_1)\geq P(y_{m2}=1|x_m,x_{m1}=x_1)\Leftrightarrow x_{i2}'\beta_0\geq x_{m2}'\beta_0, \tag{2.4}$$

$$P(y_{i0}=1|x_i,x_{i1}=x_1)\leq P(y_{m0}=1|x_m,x_{m1}=x_1)\Leftrightarrow x_{i2}'\beta_0\geq x_{m2}'\beta_0,$$

[6] Khan and Tamer (2018) showed how this notion can aid in establishing identification of regression coefficients in multinomial response models.

and

$$P(y_{i1} = 1|x_i, x_{i1} = x_1) \leq P(y_{m1} = 1|x_m, x_{m1} = x_1) \Leftrightarrow x'_{i2}\beta_0 \geq x'_{m2}\beta_0.$$

Collectively, all these identification inequalities can be used to study the conditions needed for identifying β_0. We note that when fixing regressors, individual-specific regressors drop out from these identification inequalities, so this method cannot immediately point identify their coefficients. We will revisit and elaborate on this issue in Remarks 1 and 2 later on.

To establish the (point) identification of β_0 through the identification inequalities above, the following conditions are sufficient.

CS1 The data $\{(y'_i, x'_i)'\}_{i=1}^n$ are i.i.d. from a population $\mathcal{P}$, where $y_i \equiv (y_{i0}, y_{i1}, y_{i2})'$.

CS2 (i) $\epsilon_i \perp x_i$ for all $i = 1, ..., n$, and (ii) the joint distribution of ϵ_i is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$.

CS3 For any pair of (i, m), denote $x_{imj} = x_{ij} - x_{mj}$ for $j = 1, 2$. Then, (i) without loss of generality (w.l.o.g.), $x^{(1)}_{im1}$ ($x^{(1)}_{im2}$) has almost everywhere (a.e.) positive Lebesgue density on $\mathbb{R}$ conditional on $\tilde{x}_{im1}$ ($\tilde{x}_{im2}$) and conditional on x_{im2} (x_{im1}) in a neighborhood of x_{im2} (x_{im1}) near zero, and (ii) the support of x_{im1} (x_{im2}) conditional on x_{im2} (x_{im1}) in a neighborhood of x_{im2} (x_{im1}) near zero is not contained in any proper linear subspace of $\mathbb{R}^p$.

CS4 $\beta_0 \in \text{int}(\mathcal{B})$ with $\mathcal{B} \equiv \{b \in \mathbb{R}^p || b^{(1)}| = 1\} \cap \Xi$, where $\Xi \subset \mathbb{R}^p$ is a compact set.

Assumptions CS1 and CS2 are sufficient to establish the identification inequalities like (2.3) and (2.4). Note that Assumption CS2 allows arbitrary correlation among $(\epsilon_{i1}, \epsilon_{i2})$, and our matching-based approach intrinsically accommodates flexible dependence of ϵ_i on individual specific characteristics (e.g., inter-personal heteroskedasticity). Assumption CS3(i) is a standard restriction analogous to that assumed in Manski (1975, 1985) and Han (1987), which secures the point identification, as opposed to a set identification. Assumption CS3(ii) is the familiar full-rank condition. Assumption CS4 is about scale normalization and the parameter space. As usual in discrete choice models, β_0 can only be identified up to scale. Following a substantial literature, we normalize the first element of β_0 to have absolute value one.

Our identification result for the cross-sectional multinomial response model is stated in the following theorem, which is proved in Appendix A.

Theorem 2.1. *If Assumptions CS1–CS4 hold, β_0 is identified in the parameter space $\mathcal{B}$.*

The local monotonicity in (2.3) translates into an estimation procedure, which will converge to an informative region even when all regressors have discrete support.[7] For example, given a

[7]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.

random sample of n observations, we propose the following weighted[8] rank correlation estimator, analogous to the maximum rank correlation (MRC) estimator proposed in Han (1987), defined as the maximizer, over the parameter space $\mathcal{B}$, of the objective function

$$G_{1n}(b) = \frac{1}{n(n-1)} \sum_{i \neq m} \mathbf{1}[x_{i2} = x_{m2}](y_{i1} - y_{m1}) \cdot sgn((x_{i1} - x_{m1})'b), \tag{2.5}$$

where $sgn(\cdot)$ in expression (2.5) denotes the sign function. The objective function (2.5) is associated to identification inequality (2.3) for y_{i1}. As alluded to, we can also work with y_{i0} and y_{i2}. In addition to these, by matching x_{i1} and x_{m1}, we can also construct a similar objective function motivated by (2.4)

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

It will be clear that any one or a combination of objective functions of the form above can be used for inference on β_0. To ease exposition, our discussion in the rest of this section will focus on objective function (2.5). The results can be generalized with straightforward modification.

Remark 1. *When the model contains both alternative-specific and individual-specific regressors, for example $y^*_{ij} = x'_{ij}\beta_0 + w'_i\eta_{0j} - \epsilon_{ij}$ for $j = 1, 2$ with w_i collecting all individual-specific regressors, the objective function (2.5) can only get us identification information about β_0 since with the matching $\{x_{i2} = x_{m2}, w_i = w_m\}$, $w'_i\eta_{01}$ drops out from the monotone index model. But in a special case where $\eta_{01} = \eta_{02} = \eta_0$, a two-step procedure is possible to establish identification of both β_0 and η_0. In the first step, we use objective function (2.5) to get β_0 and hence the indices $x'_{i1}\beta_0$ and $x'_{i2}\beta_0$. In the second step, by conditioning on $\{x'_{i1}\beta_0 = x'_{i2}\beta_0\}$, the probability of choosing alternative 1, for example, becomes*

$$P(y_{i1} = 1|x_i, w_i, x'_{i1}\beta_0 = x'_{i2}\beta_0) = P(x'_{i1}\beta_0 + w'_i\eta_0 - \epsilon_{i1} > 0, \epsilon_{i1} < \epsilon_{i2}).$$

This gives another version of the conditional monotone index model (monotone in $w'_i\eta_0$), and the identification of η_0 can be based on identification inequalities of the following form

$$\begin{aligned} &P(y_{i1} = 1|x_i, w_i, x'_{i1}\beta_0 = x'_{i2}\beta_0) \geq P(y_{m1} = 1|x_m, w_m, x'_{m1}\beta_0 = x'_{m2}\beta_0) \\ \Leftrightarrow &x'_{i1}\beta_0 + w'_i\eta_0 \geq x'_{m1}\beta_0 + w'_m\eta_0. \end{aligned}$$

Remark 2. *This is related to the discussion in Remark 1. For a more general case where $\eta_{01} \neq \eta_{02}$, once β_0 is "known" from the first step, identifying restrictions for $\eta_0 = (\eta'_{01}, \eta'_{02})'$ can be obtained in a second step where we condition on $\{x'_{i1}\beta_0 = x'_{m1}\beta_0 = u_1, x'_{i2}\beta_0 = x'_{m2}\beta_0 = u_2\}$ for some constants*

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

u_1 and u_2,

$$P(y_{i1}=1|x_i,w_i,x_{i1}'\beta_0=u_1,x_{i2}'\beta_0=u_2)>P(y_{m1}=1|x_m,w_m,x_{m1}'\beta_0=u_1,x_{m2}'\beta_0=u_2)$$
$$\Rightarrow\neg\{(w_i-w_m)'\eta_{01}\leq 0,(w_i-w_m)'(\eta_{01}-\eta_{02})\leq 0\}$$

with $\neg$ denoting the logical negation operator. Gao and Li (2019) study the identification and estimation of panel data multinomial response models based on identifying restrictions of this type. We believe that similar approach can be applied here to get bounds for η_0. We note that this general case would fall into the class of multiple index models. As pointed out in Lee (1995), alternative-specific regression coefficients are generally not separately point identified since identification of parameters requires that each index contains at least one distinct variable which is not contained in other indices.

Note that in the presence of continuous regressors, the probability of getting perfectly matched observations is zero.[9] Thus the value of the objective functions will always be zero. But here we can construct kernel weights as follows. To illustrate for the objective function (2.5), assuming the regressors for alternative 2 have at least one continuous component, we construct the approximate binary weight

$$K_{h_n}(x_{i2}-x_{m2})\approx\mathbf{1}[x_{i2}=x_{m2}]$$

with $K_{h_n}(\cdot)\equiv K(\cdot/h_n)$ where K is a kernel function and h_n is a bandwidth sequence that converges to 0 as $n\to\infty$. The idea is to replace the binary weights for $x_{i2}=x_{m2}$ in expression (2.5) with weights that depend inversely on the magnitude of $x_{i2}-x_{m2}$, giving more weight to observations for which $x_{i2}-x_{m2}$ is close to 0. Then we compute the estimator $\hat{\beta}$ of β_0 with the following kernel weighted objective function

$$G_{1n}^K(b)=\frac{1}{n(n-1)}\sum_{i\neq m}K_{h_n}(x_{i2}-x_{m2})(y_{i1}-y_{m1})\cdot sgn((x_{i1}-x_{m1})'b). \tag{2.6}$$

The rest of this section concerns the asymptotic properties of the estimator $\hat{\beta}$ defined as the maximizer of objective function (2.6).[10] Before presenting additional regularity conditions and the main results, we introduce some new notations to ease exposition:

- $f_2(\cdot)$ denotes the PDF of x_{im2}. $F_x(\cdot)$ ($f_x(\cdot)$) denotes the joint probability distribution (density) function of x_i. $f_{x_j}(\cdot)$ denotes the marginal PDF of x_{ij} for $j=1,2$. $f_{x_2|x_1}(\cdot)$ denotes the

[9] But note that this depends on the choice in question. For example, consider the same 3-alternative setting. Suppose for alternative 1, the regressor vector has one continuous component with support on the real line, but its other components are discrete. Suppose for alternative 2, all the components of the regressor vector are discrete. Then we can match as in (2.5), and this in fact will point identify β_0.

[10] To streamline exposition, we will focus on the case where all components of x_{i2} are continuous. In practice, when the regressor vector contains both continuous and discrete components, one could apply the kernel weight to the former and binary weight to the latter.

conditional PDF of x_{i2} conditional on x_{i1}.

- $y_{im1} \equiv y_{i1} - y_{m1}$ and $q_{im}(b) \equiv y_{im1} \cdot sgn(x'_{im1}b)$.
- $B(x_{i1}, x_{m1}, x_{i2}, x_{m2}) \equiv E[y_{im1}|x_{i1}, x_{m1}, x_{i2}, x_{m2}]$ and $S_{im}(b) \equiv sgn(x'_{im1}b)$.
- $\tau_i(b) \equiv E[q_{im}(b)|x_{i1}, x_{i2}, x_{m2} = x_{i2}] = \int B(x_{i1}, x_{m1}, x_{i2}, x_{i2})S_{im}(b)f_x(x_{m1}, x_{i2})dx_{m1}$. We use $\nabla_1\tau_i(b)$ and $\nabla_2\tau_i(b)$ to denote the gradient and Hessian matrix of function $\tau_i(\cdot)$ evaluated at b, respectively.

We impose the following regularity conditions:

CS5 $f_2(\cdot)$ is absolutely continuous, bounded from above on its support, strictly positive in a neighborhood of zero, and continuously differentiable with bounded first derivatives.

CS6 For all $b \in \mathcal{B}$, $E[q_{im}(b)|x_{im2} = \cdot]$ is continuously differentiable with bounded first derivatives. $B(\cdot, \cdot, \cdot, \cdot)$ is κ_B^{th} continuously differentiable with bounded κ_B^{th} derivatives, and $f_{x_2|x_1}(\cdot)$ is κ_f^{th} continuously differentiable with bounded κ_f^{th} derivatives. Denote $\kappa = \kappa_B + \kappa_f$. κ is an even integer greater than p.

CS7 Let $\|\cdot\|_F$ denote the Frobenius norm, $\mathcal{X}$ denote the support of x_i, and $\mathcal{N}$ denote a neighborhood of β_0. Then, for all $x_i \in \mathcal{X}$ and $b \in \mathcal{N}$,

- There exists an integrable function $\phi_s(\cdot)$ such that $\int |S_{im}(b) - S_{im}(\beta_0)|f_{x_1}(x_{m1})dx_{m1} \leq \phi_s(x_{i1})\|b - \beta_0\|$.
- All mixed second partial derivatives of $\tau_i(b)$ exist on $\mathcal{N}$.
- There is an integrable function $\phi_\tau(\cdot)$ such that $\|\nabla_2\tau_i(b) - \nabla_2\tau_i(\beta_0)\|_F \leq \phi_\tau(x_i)\|b - \beta_0\|$.
- $E[\|\nabla_1\tau_i(\beta_0)\|^2] < \infty$, $E[\|\nabla_2\tau_i(\beta_0)\|_F] < \infty$, and $E[\nabla_2\tau_i(\beta_0)]$ is negative definite.

CS8 The function $K : \mathbb{R}^p \mapsto \mathbb{R}$ used to construct the weight in (2.6) is an κ^{th} order bias-reducing kernel. $K(\cdot)$ is continuously differentiable and also assumed to satisfy the following conditions: (i) $\sup_{v\in\mathbb{R}^p} |K(v)| < \infty$, (ii) $\int K(v)dv = 1$, (iii) $\int |v|_1|K(v)|dv < \infty$, where $|\cdot|_1$ denotes the l_1-norm, and (iv) for positive integers $\iota_1, ..., \iota_p$ satisfying $0 < \iota_1 + \cdots + \iota_p \leq \kappa$,

$$\int v_1^{\iota_1} v_2^{\iota_2} \cdots v_p^{\iota_p} K(v_1, ..., v_p)dv_1 \cdots dv_p \begin{cases} = 0 & \text{if } \iota_1 + \cdots + \iota_p < \kappa \\ \neq 0 & \text{if } \iota_1 + \cdots + \iota_p = \kappa \end{cases}.$$

CS9 The bandwidth sequence h_n used to construct the kernel weight in (2.6) is a sequence of positive numbers such that as $n \to \infty$: (i) $h_n \to 0$, (ii) $\sqrt{n}h_n^\kappa \to 0$, and (iii) $\sqrt{n}h_n^p \to \infty$.

The boundedness and smoothness restrictions placed in Assumptions CS5 and CS6 are needed for proving the uniform convergence of the objective function to its population analogue and deriving the root-n rate of $\hat{\beta}$. Assumption CS7 is analogous to the regularity conditions imposed in

Sherman (1993). Assumptions CS8 and CS9 place mild restrictions on kernel functions and tuning parameters, all of which are standard in the literature.

The theorem below establishes the $\sqrt{n}$-consistency and asymptotic normality of the proposed estimator. The proof, as presented in Appendix A, follows from similar arguments to those used in Han (1987), Sherman (1993, 1994a,b), and Abrevaya, Hausman, and Khan (2010).

Theorem 2.2. *If Assumptions CS1–CS9 hold, then (i)* $\hat{\beta} \xrightarrow{p} \beta_0$*, and (ii)* $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, V^{-1}\Lambda V^{-1})$*, where* $V = E[\nabla_2 \tau_i(\beta_0)]$ *and* $\Lambda = 4E[\nabla_1 \tau_i(\beta_0)\nabla_1 \tau_i(\beta_0)']$.

Theorem 2.2 indicates that our rank estimator $\hat{\beta}$ is asymptotically normal and has asymptotic variance of regular "sandwich" structure. To make inference, Sherman (1993) proposes to use the numerical derivative method of Pakes and Pollard (1989) to estimate the moments V and Λ in the asymptotic variance. Hong, Mahajan, and Nekipelov (2015) study the application of the numerical derivative method in a wide range of extremum estimators, including second-order U-statistics. Cavanagh and Sherman (1998) suggest estimating V and Λ nonparametrically. These methods, however, require selecting additional tuning parameters, and hence are hard to implement. Subbotin (2007) shows that the asymptotic variance of the MRC estimator can be consistently estimated by the nonparametric bootstrap under mild conditions, which makes the bootstrap a potentially attractive method for carrying out inference in various empirical studies. See also Jin, Ying, and Wei (2001) for an alternative resampling method by perturbing the objective function repeatedly.

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 unobserved heterogeneity across both individuals and alternatives. 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 Rasch (1960) and Andersen (1970). For a review of the early work on this model, see Chamberlain (1984) (Section 3), and for a survey of more recent contributions, see Arellano and Honoré (2001) (Sections 4–9). 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), Graham and Powell (2012), Hoderlein and White (2012), Chernozhukov, Fernández-Val, Hahn, and Newey (2013), Chen, Khan, and Tang (2015), and Khan, Ponomareva, and Tamer (2016), to name a few.

In this section of the paper, we consider a panel data model for multinomial response as in Chamberlain (1980) and Chamberlain (1984) where the indirect utility and observed choices can be expressed as

$$\begin{aligned} y^*_{i0t} &= 0, \\ y^*_{ijt} &= x'_{ijt}\beta_0 + \alpha_{ij} - \epsilon_{ijt}, \end{aligned}$$

and

$$y_{ijt} = \mathbf{1}[y^*_{ijt} > y^*_{ikt}, \forall k \neq j]$$

for $i = 1, ..., n$, $j, k \in \{0, 1, ..., J\}$, and $t = 1, ..., T$. In our notation, for the subscript ijt, the first component i denotes the individual, the second component j denotes the alternative, and the third component t denotes the time period. As in the cross-section case, we impose the (location) normalization that $y^*_{i0t} = 0$ for all $t = 1, ..., T$. Note that the random utilities specified above include a set of *fixed effects* α_{ij} that are *both* individual and alternative specific. Throughout, no assumptions are made on the distribution of $\alpha_i \equiv (\alpha_{i1}, ..., \alpha_{iJ})'$ conditional on $x_i \equiv (x'_{i11}, ..., x'_{iJ1}, ..., x'_{i1T}, ..., x'_{iJT})'$ and $\epsilon_i \equiv (\epsilon_{i11}, ..., \epsilon_{iJ1}, ..., \epsilon_{i1T}, ..., \epsilon_{iJT})'$.

Here we consider identification and asymptotics with J, T fixed and $n \to \infty$. Existing results for panel data binary choice models with fixed effects include Rasch (1960), Andersen (1970), Manski (1987), and Chamberlain (2010), among others. The literature on multinomial choice models for panel data is more limited. The conditional likelihood method proposed by Chamberlain (1980) is consistent and $\sqrt{n}$-normal for the Logit specification. Recent semiparametric results include Pakes and Porter (2014) and Shi, Shum, and Song (2018). The former is concerned with partial identification, while the latter achieves point identification. Our work is in line with Manski (1987), Pakes and Porter (2014) and Shi, Shum, and Song (2018) in the sense that our identification strategy relies on similar *group homogeneity* conditions as ones adopted by the aforementioned papers.

Specifically, letting $\epsilon_{it} \equiv (\epsilon_{i1t}, ..., \epsilon_{iJt})'$ and $x_{it} \equiv (x'_{i1t}, ..., x'_{iJt})'$, we assume that for all $s \neq t$, $\epsilon_{is} \overset{d}{=} \epsilon_{it}|(\alpha_i, x_{is}, x_{it})$. To ease exposition, our results for this section will be presented by a model with $J = 2$ an $T = 2$.[11] Our approach can be modified in a straightforward manner to be applied to data with more alternatives or longer panel.

[11] So the choice set is $\{0, 1, 2\}$, and we use a single pair of time periods, 1 and 2.

Assuming $\epsilon_{i1} \overset{d}{=} \epsilon_{i2}|(\alpha_i, x_{i1}, x_{i2})$, we have the following conditional monotone index model

$$P(y_{i11} = 1|x_i, x_{i21} = x_2) \geq P(y_{i12} = 1|x_i, x_{i22} = x_2) \Leftrightarrow x'_{i11}\beta_0 \geq x'_{i12}\beta_0, \tag{3.1}$$

which is the key for our identification results. The identification inequality (3.1) is analogous to (2.3) for the cross-sectional case, but now we match and do 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 at least one alternative are time-varying, and whose regressors for other alternatives have overlapping support over time.

Remark 3. *Note that $P(y_{i0t} = 1|x_i, x_{i2t} = x_2)$ and $P(y_{i2t} = 1|x_i, x_{i2t} = x_2)$ are both decreasing in $x'_{i1t}\beta_0$, which gives additional identification inequalities for β_0*

$$P(y_{i01} = 1|x_i, x_{i21} = x_2) \leq P(y_{i02} = 1|x_i, x_{i22} = x_2) \Leftrightarrow x'_{i11}\beta_0 \geq x'_{i12}\beta_0$$

and

$$P(y_{i21} = 1|x_i, x_{i21} = x_2) \leq P(y_{i22} = 1|x_i, x_{i22} = x_2) \Leftrightarrow x'_{i11}\beta_0 \geq x'_{i12}\beta_0.$$

Furthermore, one can use analogous arguments to deduce

$$P(y_{i21} = 1|x_i, x_{i11} = x_1) \geq P(y_{i22} = 1|x_i, x_{i12} = x_1) \Leftrightarrow x'_{i21}\beta_0 \geq x'_{i22}\beta_0,$$

$$P(y_{i01} = 1|x_i, x_{i11} = x_1) \leq P(y_{i02} = 1|x_i, x_{i12} = x_1) \Leftrightarrow x'_{i21}\beta_0 \geq x'_{i22}\beta_0,$$

and

$$P(y_{i11} = 1|x_i, x_{i11} = x_1) \leq P(y_{i12} = 1|x_i, x_{i12} = x_1) \Leftrightarrow x'_{i21}\beta_0 \geq x'_{i22}\beta_0.$$

These inequalities contain identification information about β_0.

The monotonic relation established in (3.1) motivates one objective function we work with[12]

$$G_n^{SP}(b) = \frac{1}{n}\sum_i \mathbf{1}[x_{i21} = x_{i22}](y_{i11} - y_{i12}) \cdot sgn\left((x_{i11} - x_{i12})'b\right). \tag{3.2}$$

Given a random sample of individuals $i = 1, ..., n$, the estimator of β_0 is defined as the maximizer, over a parameter space $\mathcal{B}$, of (3.2).

Note that objective function (3.2) 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

[12] In addition to (3.2), similar objective functions can be constructed using identification inequalities provided in Remark 3. Collectively, any one or a combination of these objective functions can be used for estimating β_0. Furthermore, this objective function can be naturally modified for the case when there are more time periods. This is demonstrated in the empirical example in Section 5.

switchers, or individuals whose choice changes over time.

For the case where x_{i21} and x_{i22} contain continuous components, we replace the indicator function in (3.2) with a kernel function to yield

$$G_n^{SP,K}(b) = \frac{1}{n}\sum_i K_{h_n}(x_{i21} - x_{i22})(y_{i11} - y_{i12})sgn\left((x_{i11} - x_{i12})'b\right). \tag{3.3}$$

with $K_{h_n}(\cdot) \equiv K(\cdot/h_n)$, 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 $\hat{\beta}$ of objective function (3.3) is a consistent estimator.[13] To facilitate exposition in stating our conditions, we first introduce the following notations.

Notation: To lighten the notation, we will suppress the subscript i in the rest of this section whenever it is clear that all variables are for each individual. Let $y_{it} \equiv (y_{i0t}, y_{i1t}, y_{i2t})'$ for $t = 1, 2$ and $y_i \equiv (y_{i1}', y_{i2}')'$. For generic random vectors v_{js} and v_{jt}, $v_{j(st)} \equiv v_{js} - v_{jt}$, e.g., $x_{1(12)} = x_{11} - x_{12}$. Denote $\rho(b) = y_{1(12)} \cdot sgn(x_{1(12)}'b)$ for all $b \in \mathbb{R}^p$.

Next, we outline the regularity conditions for point identification and consistency of our semiparametric estimator based on the objective function (3.3).

SP1 $\{(y_i, x_i)\}_{i=1}^n$ is a random sample from a population $\mathcal{P}$.

SP2 $\beta_0 \in \text{int}(\mathcal{B})$, where $\mathcal{B} = \{b \in \mathbb{R}^p : \|b\| = 1, b^{(1)} \neq 0\}$.[14]

SP3 (i) $\epsilon_1 \overset{d}{=} \epsilon_2 | (\alpha, x)$, (ii) $\epsilon_t | (\alpha, x)$, $t = 1, 2$, has absolutely continuous distribution on $\mathbb{R}^2$.

SP4 $x_{1(12)}^{(1)}$ w.l.o.g. has a.e. positive Lebesgue density conditional on $\tilde{x}_{1(12)}$ and conditional on $x_{2(12)}$ in a neighborhood of $x_{2(12)}$ near zero.

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 $f_{x_{2(12)}}(\cdot)$ that is bounded from above on its support and strictly positive in a neighborhood of zero.[15]

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

[13] As was the case in the cross-sectional model, point identification is not attainable when all the regressors are discrete, but objective function (3.3) is still useful for obtaining informative bounds for β_0.

[14] The scale normalization here follows the convention adopted by a substantial literature. See e.g., Manski (1987) and Kim and Pollard (1990). An alternative way for scale normalization is to assume w.l.o.g. that the first element of b has absolute value one.

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

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)$\int |v|_1|K(v)|dv < \infty$, where $|\cdot|_1$ denotes the l_1-norm.

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

Theorem 3.1. *Suppose Assumptions SP1–SP9 hold. Then, (i) β_0 is point identified relative to all $b \in \mathcal{B}\backslash\{\beta_0\}$, and (ii) $\hat{\beta} \xrightarrow{p} \beta_0$, where $\hat{\beta}$ is a sequence of the solutions to the problem* $\max_{b\in\mathcal{B}} G_n^{SP,K}(b)$.

Next, we derive the rate of convergence and asymptotic distribution of $\hat{\beta}$. To examine the effect of dimensionality in the number of alternatives, we consider the general case with $T = 2$ and $J + 1$ alternatives (numbered from 0 to J, $J \geq 2$). For notational 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

$$\frac{1}{n}\sum_i K_{h_n}(z_{i1})z_{i2} \cdot sgn(z'_{i3}b).$$

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 $f_{z_1|z_2\neq 0,z_3}(\cdot)$ 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.

SP8' $K : \mathbb{R}^{(J-1)p} \mapsto \mathbb{R}$ is a kernel density function of bounded variation and bounded support that satisfies: (i) $\sup_{v\in\mathbb{R}^{(J-1)p}} |K(v)| < \infty$, (ii) $\int K(v)dv = 1$, and (iii)$\int \|v\|^2|K(v)|dv < \infty$.

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 and asymptotic distribution 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 $\hat{\beta}$ be a sequence of the solutions to the problem*

$$\max_{b\in\mathcal{B}} \frac{1}{n}\sum_i K_{h_n}(z_{i1})z_{i2} \cdot sgn(z'_{i3}b).$$

Then, (i) $\hat{\beta} - \beta_0 = O_p((nh_n^{(J-1)p})^{-1/3})$*, and (ii)*

$$(nh_n^{(J-1)p})^{1/3}(\hat{\beta} - \beta_0) \xrightarrow{d} \arg\max_{\mathbf{s}\in\mathbb{R}^p} Z(\mathbf{s}),$$

where $Z(\mathbf{s})$ *is a Gaussian process with continuous sample paths, expected value* $\mathbf{s}'V\mathbf{s}/2$*, and covariance kernel* $H(\mathbf{s}_1, \mathbf{s}_2)$ *for* $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^p$*.* V *and* $H(\cdot,\cdot)$ *are defined in expressions (B.8) and (B.13), respectively.*

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 settings, rate of convergence depends on the number of alternatives, J. As with more alternatives, we are matching more covariates. Proofs of the above results are collected in Appendix B.

The asymptotic distribution of $\hat{\beta}$ does not have an analytic form, making inference difficult to conduct. One may consider a smoothed MS approach (e.g., Horowitz (1992), Kyriazidou (1997), and Charlier (1997)), which has the potential to yield an asymptotically normal estimator. However, smoothing the objective function involves choosing additional kernel functions and tuning parameters. As an alternative, we recommend to use bootstrap-based procedures for inference. As shown in Abrevaya and Huang (2005), the classic bootstrap is inconsistent for the MS estimators, and hence we expect that the classic bootstrap does not work for our estimators, either. For the ordinary MS estimator, valid inference can be conducted using subsampling (Delgado, Rodríguez-Poo, and Wolf (2001)), m-out-of-n bootstrap (Lee and Pun (2006)), the numerical bootstrap (Hong and Li (2020)), and a model-based bootstrap procedure that analytically modifies the criterion function (Cattaneo, Jansson, and Nagasawa (2020)), among other procedures. Ouyang and Yang (2020a,b) show that Hong and Li's (2020) and Cattaneo et al.'s (2020) methods, with certain modifications, can be justified to be valid for kernel weighted MS estimators of similar structure. We expect the same methods apply to the estimator of this paper.

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, coefficients on the lagged dependent variables.

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 the same choice in the future as compared to another individual who has not experienced the event or made that choice. They 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. So having experienced 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. This explanation revolves around heterogeneity (rather than state dependence): some individuals are more likely to make a specific choice than others due to unobserved factors. The econometrics literature on the topic has provided various models to 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 multinomial response models with 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 (i.e., $n \to \infty$ and $T < \infty$). To the best of our knowledge, this is the first semiparametric (distribution-free) approach in the literature to study this problem.

Our results focus on point identification. As in Section 3.1, we illustrate our approach with $J = 2$. A particular model that we consider can be expressed as follows.

$$\begin{aligned} y^*_{i0t} &= 0, \\ y^*_{i1t} &= x'_{i1t}\beta_0 + \gamma_0 y_{i1t-1} + \alpha_{i1} - \epsilon_{i1t}, \\ y^*_{i2t} &= x'_{i2t}\beta_0 + \alpha_{i2} - \epsilon_{i1t}, \end{aligned}$$

and $y_{ijt} = \mathbf{1}[y^*_{ijt} > y^*_{ikt}, \forall k \neq j]$ for $t = 1, 2, ..., T$. Following the literature, we define period 0 as the initial period, and assume that $y_{i0} \equiv (y_{i00}, y_{i10}, y_{i20})'$ are observed, although the model is not specified in the initial period. Throughout this section, we focus on the case with $T = 3$, the minimum T required for applying our identification approach.

In this model, the parameters of interest are $\theta_0 \equiv (\beta'_0, \gamma_0)'$. 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 three time periods ($T = 2$).[16] Honoré and Kyriazidou (2000) show point identification[17] of β_0 and γ_0 when there are four time periods ($T = 3$). Their identification is based on conditioning on the subset of the population whose regressors do not change in periods 2 and 3. Finally, Khan, Ponomareva, and Tamer (2019) derive sharp bounds for preference parameters in dynamic binary choice models with fixed effects under weak conditions (allowing for time trends, time dummies, etc.).

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

[17] 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.

Our identification strategy for the dynamic multinomial response model is based on conditioning on the sub-population whose regressors are time-invariant in different manners, depending on which alternative they are associated with. Specifically, in the three alternatives, four periods setting above, we condition on the sub-population whose regressor values for alternative 2 do not change in periods 1, 2, and 3 and whose regressor values for choice 1 do not change over time in periods 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 all regressors are discrete, the estimator will converge at the parametric rate with an asymptotically normal distribution, so conducting inference is relatively easy. We demonstrate both parametric (Logit) and semiparametric methods for the dynamic multinomial response model considered in order.

For the dynamic multinomial Logit model, we consider the following conditional likelihood function:[18]

$$\sum_i \mathbf{1}[x_{i21} = x_{i22} = x_{i23}, x_{i12} = x_{i13}]\mathbf{1}[y_{i11} \neq y_{i12}] \cdot \log\left(\frac{\exp\left((x_{i11} - x_{i12})'b + r(y_{i10} - y_{i13})\right)^{y_{i11}}}{1 + \exp\left((x_{i11} - x_{i12})'b + r(y_{i10} - y_{i13})\right)}\right).$$

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 and 2. For general J and T, the number of such sequences is $(J+1)^{T+1} - (J+1)^3$, then coding the estimator may be cumbersome, especially for cases with large J or large T.[19] Our estimator differs from theirs, as here we effectively transform a multinomial response problem to a binary choice problem through additionally matching x_{i21} and x_{i22}, 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 choice model.

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

$$\frac{1}{n}\sum_i \mathbf{1}[x_{i21} = x_{i22} = x_{i23}, x_{i12} = x_{i13}](y_{i11} - y_{i12}) \cdot sgn\left((x_{i11} - x_{i12})'b + r(y_{i10} - y_{i13})\right).$$

Note that for point identification, we require that at least one of the components of the regressors

[18] Throughout this section, we deliberately keep the notation as close as possible to Honoré and Kyriazidou (2000).
[19] For example, in our empirical illustration, we have $J = 3$ and $\max_i T_i = 77$ (unbalanced panel).

for alternative 1 to be continuously distributed on a large support. Consequently, when matching regressors for this choice, we would need to assign kernel weights as illustrated in previous sections. Denote $\theta = (b', r)'$, $z_{i1} = (x'_{i2(12)}, x'_{i2(23)}, x'_{i1(23)})'$, $z_{i2} = y_{i1(12)}$, and $z_{i3} = (x'_{i1(12)}, y_{i1(03)})'$. In practice, we work with the objective function

$$G_n^{DP,K}(\theta) = \frac{1}{n}\sum_i K_{h_n}(z_{i1}) z_{i2} \cdot sgn(z'_{i3}\theta). \tag{3.4}$$

Under the standard "initial conditions" assumption as in e.g., Honoré and Kyriazidou (2000),[20] the maximizer $\hat{\theta}$ of this objective function can be shown to be consistent, although as in the static model, the limiting distribution is nonstandard.

Remark 4. *The objective function (3.4) is constructed based on the following identification inequality, which is proved in Appendix B.*

$$P(A|x_i, \alpha_i, \Omega) \geq P(B|x_i, \alpha_i, \Omega) \Leftrightarrow x'_{i11}\beta_0 + \gamma_0 y_{i10} \geq x'_{i12}\beta_0 + \gamma_0 y_{i13},$$

where events $\Omega \equiv \{x_{i21} = x_{i22} = x_{i23}, x_{i12} = x_{i13}\}$, $A \equiv \{y_{i10} = d_0, y_{i11} = 1, y_{i12} = 0, y_{i13} = d_3\}$, *and* $B \equiv \{y_{i10} = d_0, y_{i11} = 0, y_{i12} = 1, y_{i13} = d_3\}$ *for* $(d_0, d_3) \in \{0,1\}^2$. *The key idea, as outlined above, is to turn the multinomial response model into a binary choice model by matching the covariates for all but one inside alternatives in different time periods and use data on "switchers".*

We next present conditions that are sufficient for consistency and asymptotic distribution of the proposed estimator $\hat{\theta}$. As in Section 3.1, we suppress the subscript i to lighten the notation. Denote $\psi(\theta) = z_2 \cdot sgn(z'_3\theta)$ for all $\theta \in \mathbb{R}^{p+1}$. We assume:

DP1 $\{(y_i, x_i)\}_{i=1}^n$ is a random sample from a population $\mathcal{P}$.

DP2 $\theta_0 \in \text{int}(\Theta)$, where $\Theta = \{\theta = (b', r)' \in \mathbb{R}^{p+1} : \|\theta\| = 1, b^{(1)} \neq 0\}$.

DP3 For almost all (x, α), (i) $\epsilon_t \perp (x, y_0)|\alpha$ holds for all $t = 1, 2, 3$ and (ii) $\epsilon_t|\alpha$ is i.i.d. over time[21] having absolutely continuous distribution on $\mathbb{R}^2$.

DP4 $z_3^{(1)}$ w.l.o.g. has a.e. positive Lebesgue density conditional on $\tilde{z}_3$ and conditional on z_1 in a neighborhood of z_1 near zero.

DP5 The support of z_3 conditional on z_1 in a neighborhood of z_1 near zero is not contained in any proper linear subspace of $\mathbb{R}^p$.

[20] Specifically, for the model at hand, the initial conditions assumption would be that $P(y_{ij0} = 1|x_i, \alpha_i) = p_{ij0}(x_i, \alpha_i)$ for $j = 0, 1, 2$, where the functional form of $p_{ij0}(\cdot, \cdot)$ is left unspecified. With this assumption, we do not require specifying the model in the initial period, since the value of the dependent variable is not assumed to be known in periods prior to the sample. It is worth noting that, taken together, $p_{ij0}(\cdot, \cdot)$ and $F_{\alpha_i|x_i}$ give the joint distribution of $(y_{i00}, y_{i10}, y_{i20}, \alpha_i)'$.

[21] Note that it is possible to generalize the results in this section to allow the distribution of $\epsilon_t|\alpha$ to vary across individuals, provided that it does not differ over time for a given individual.

DP6 $z_1 \in \mathbb{R}^{3p}$ is absolutely continuously distributed with bounded density $f_{z_1}(\cdot)$. Both $f_{z_1}(\cdot)$ and the conditional density $f_{z_1|z_2\neq 0,z_3}(\cdot)$ are strictly positive in a neighborhood of zero.

DP7 For all $\theta \in \Theta$, $f_{z_1}(\cdot)$ and $E[\psi(\theta)|z_1 = \cdot]$ are twice differentiable on their support with bounded second-order derivatives.

DP8 $K : \mathbb{R}^{3p} \mapsto \mathbb{R}$ is a kernel density function of bounded variation and bounded support that satisfies: (i) $\sup_{v\in\mathbb{R}^{3p}} |K(v)| < \infty$, (ii) $\int K(v)dv = 1$, and (iii)$\int \|v\|^2|K(v)|dv < \infty$.

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

The above conditions suffice for point identification and asymptotic properties of our proposed estimator as stated in the following theorem, proved in Appendix B.

Theorem 3.3. *Suppose Assumptions DP1–DP9 hold. Then, (i) θ_0 is identified relative to all $\theta \in \Theta \setminus \{\theta_0\}$, (ii) $\hat{\theta} \xrightarrow{p} \theta_0$, (iii) $\hat{\theta} - \theta_0 = O_p((nh_n^{3p})^{-1/3})$, and (iv)*

$$(nh_n^{3p})^{1/3}(\hat{\theta} - \theta_0) \xrightarrow{d} \arg\max_{\mathbf{s}\in\mathbb{R}^{p+1}} Z(\mathbf{s}),$$

where $Z(\mathbf{s})$ is a Gaussian process with continuous sample paths, expected value $\mathbf{s}'V\mathbf{s}/2$, and covariance kernel $H(\mathbf{s}_1, \mathbf{s}_2)$ for $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^{p+1}$. V and $H(\cdot,\cdot)$ are defined in expressions (B.8) and (B.13), respectively.

Remark 5. *The conclusions of Theorem 3.3 can be generalized to general cases with $J+1$ alternatives. Particularly, with straightforward adjustments to Assumptions DP1–DP9, we can conclude that $\hat{\theta} - \theta_0 = O_p((nh_n^{(2J-1)p})^{-1/3})$ and $(nh_n^{(2J-1)p})^{1/3}(\hat{\theta} - \theta_0) \xrightarrow{d} \arg\max_{\mathbf{s}\in\mathbb{R}^{p+1}} Z(\mathbf{s})$ where $Z(\cdot)$ is of the same form as defined in Theorem 3.3. For inference, the discussion on the use of the bootstrap after Theorem 3.2 applies here as the estimator has essentially the same structure.*

3.2.1 Identification with More General Feedback Effects

The identification approach described in Remark 4 extends to dynamic model with more general "feedback" effects:

$$\begin{aligned} y^*_{i0t} &= 0, \\ y^*_{ijt} &= x'_{ijt}\beta_0 + \gamma_{0,j1}y_{i1t-1} + \gamma_{0,j2}y_{i2t-1} + \alpha_{ij} - \epsilon_{ijt},\ j = 1, 2, \end{aligned}$$

where $\gamma_{0,jk}$ is the feedback effect when a choice of alternative k at $t-1$ is followed by choice j at time t, where $j, k \in \{1, 2\}$. Note that due to location normalization and multicollinearity, $\gamma_{0,0j}$ and $\gamma_{0,j0}$ for all $j = 1, 2$ are not identified[22] (or equivalently, we impose normalization $\gamma_{0,0j} = \gamma_{0,j0} = 0$ for all

[22]This is similar to the dynamic multinomial Logit model considered in Magnac (1997).

$j = 1, 2$). If we further assume all "cross" feedback parameters are zero, i.e., $\gamma_{0,jk} = 0$ for all $j \neq k$, then $(\beta_0', \gamma_{0,11}, \gamma_{0,22})'$ can be identified based on identification inequalities similar to that presented in Remark 4. For example, letting $\mathcal{E} = \{x_{i21} = x_{i22} = x_{i23}, x_{i12} = x_{i13}, y_{i20} = y_{i21} = y_{i22} = 0\}$, we have

$$P(A|x_i, \alpha_i, \mathcal{E}) \geq P(B|x_i, \alpha_i, \mathcal{E}) \Leftrightarrow x_{i11}'\beta_0 + \gamma_{0,11}y_{i10} \geq x_{i12}'\beta_0 + \gamma_{0,11}y_{i13},$$

for all $d_0 \neq d_3$. As pointed out by a referee, "No cross feedback" assumption is restrictive and may not be intuitively plausible when alternatives are close substitutes. From this point of view, this specification may be more reasonable for the case where alternatives are not very similar to each other. For the model with unrestricted cross feedback effects (with exception of the normalization imposed on $\gamma_{0,0j}$ and $\gamma_{0,j0}$), it is known in the literature (see Section 4.3 of Honoré and Kyriazidou (2000)) that the parameters are identified with a multinomial Logit specification, heavily relying on distributional assumptions leading to a logistic distribution (with the IIA property). When no distributional restrictions are imposed, we can still point identify β_0, for example, via

$$P(A|x_i, \alpha_i, \mathcal{E}) \geq P(B|x_i, \alpha_i, \mathcal{E}) \Leftrightarrow x_{i11}'\beta_0 \geq x_{i12}'\beta_0,$$

for all $d_0 = d_3$. Bounds for feedback parameters can be obtained by identification restrictions constructed using "logical contradiction" as in Remark 2. As an illustrating example, consider event $\Upsilon = \{x_{i12} = x_{i13}, x_{i22} = x_{i23}, y_{i20} = y_{i21} = y_{i22} = 0\}$. Then we can write

$$\begin{aligned} &P(A|x_i, \alpha_i, \Upsilon) > P(B|x_i, \alpha_i, \Upsilon) \\ \Rightarrow &\neg\{x_{i11}'\beta_0 + \gamma_{0,11}y_{i10} \leq x_{i12}'\beta_0 + \gamma_{0,11}y_{i13}, x_{i21}'\beta_0 + \gamma_{0,21}y_{i10} \geq x_{i22}'\beta_0 + \gamma_{0,21}y_{i13}\}, \end{aligned}$$

for all $d_0 \neq d_3$. However, a detailed analysis of (partial) identification based on inequalities of this type is beyond the scope of this paper as it would involve a completely different identification approach than the ones introduced in this paper.

4 Simulation Study

In this section, we investigate the relative finite sample performance of the proposed estimation procedures in cross-sectional and panel data (both static and dynamic) designs. We generate 1000 replications for each of the six designs described below, using sample sizes n ranging from 250 to 10000. In all designs, the regressor vector always has one and only one component that is continuously distributed with all the rest being binary,[23] and the error vector follows a multivariate normal (MVN) distribution that allows for correlation across components.

[23] We expect that the design of the regressors in Monte Carlo studies may have a large effect on the performance of our matching-based methods. In order to investigate this issue, we also run experiments with all continuous regressors and report the results in Appendix C.

For the cross-sectional model, we generate data from three designs, varying the number of regressors and the number of alternatives in the choice set. The first two are for a model with 3 alternatives ($J = 2$), and we increase the number of regressors p from 3 to 5. This is meant to give an idea of the sensitivity of our estimator to the dimensionality of the regressor space. In the third design, we considered 3 regressors but 5 alternatives. Here, we aim to examine 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 3 alternatives, 3 regressors, and 2 time periods. The second panel data design is for the dynamic model where there are 3 alternatives and 3 regressors with the second of the two binary regressors being the lagged choice. For this model, we simulate 4 periods of data as this is the minimum length of panel required for our identification approach.

For each of these six designs and varying sample sizes, we report the mean bias (MEAN) and root mean squared error (RMSE) of the corresponding estimator. Since these statistics can be sensitive to outliers, we also present the median bias (MED) and the median absolute error (MAE). Below we state the details of each of the designs considered and the Monte Carlo results for our estimators in order.

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 indirect utility functions:

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

where $x_{ij}^{(1)}, x_{ij}^{(2)}, x_{ij}^{(3)}$ denote the 3 components of the vector x_{ij}, $\beta_1 = \beta_2 = 1$, $(x_{i1}^{(1)})_{i=1,\ldots,n}$ are independent $N(0,1)$ random variables, $(x_{i1}^{(2)}, x_{i1}^{(3)}, x_{i2}^{(1)}, x_{i2}^{(2)}, x_{i2}^{(3)})_{i=1,\ldots,n}$ are independent Bernoulli random variables with parameter 0.5, and

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

To implement our weighted rank correlation estimator, we use the sixth-order Gaussian kernel and bandwidth $h_n = n^{-1/5}$. Table 1 reports the results for this benchmark design.

Table 1: (Design 1) Cross-Sectional Design with $J = 2$ and $p = 3$

	β_1				β_2			
	MEAN	RMSE	MED	MAE	MEAN	RMSE	MED	MAE
$n = 250$	0.0256	0.2853	0.0041	0.1772	0.0258	0.2831	0.0027	0.1757
$n = 500$	0.0150	0.1858	0.0013	0.1055	0.0079	0.1820	0.0009	0.0999
$n = 1000$	0.0039	0.1192	0.0008	0.0683	0.0022	0.1170	-0.0001	0.0656

As our cross-sectional estimator is "localized" (matching covariates associated with $J-1$ alternatives), one may be worried about that the dimensionality of the design (both in the regressor space and choice space) may have a substantial effect on the simulation results. 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 adding two regressors to the benchmark design. Specifically, we consider the DGP with indirect utility functions:

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

  where $\beta_1 = \beta_2 = 1$, $\beta_3 = \beta_4 = 0$, $(x_{i1}^{(1)})_{i=1,\ldots,n}$ are independent $N(0,1)$ random variables, and $(x_{i1}^{(2)}, x_{i1}^{(3)}, x_{i1}^{(4)}, x_{i1}^{(5)}, x_{i2}^{(1)}, x_{i2}^{(2)}, x_{i2}^{(3)}, x_{i2}^{(4)}, x_{i2}^{(5)})_{i=1,\ldots,n}$ are independent Bernoulli random variables with parameter 0.5. The DGP is essentially the same as that for the benchmark design and the only difference is that two additional regressors are included in the estimation.

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

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

  where $\beta_1 = \beta_2 = 1$, $(x_{i1}^{(1)})_{i=1,\ldots,n}$ are independent $N(0,1)$ random variables, $(x_{i1}^{(2)}, x_{i1}^{(3)})_{i=1,\ldots,n}$ and $(x_{ij}^{(l)})_{i=1,\ldots,n;j=2,3,4;l=1,2,3}$ are independent Bernoulli random variables with parameter 0.5, and

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

The results of these two experiments, using the same kernel function and bandwidth as in Design 1, are summarized in Tables 2 and 3, respectively.[24]

Table 2: (Design 2) Cross-Sectional Design with $J = 2$ and $p = 5$

	β_1				β_2			
	MEAN	RMSE	MED	MAE	MEAN	RMSE	MED	MAE
$n = 250$	0.0276	0.3265	0.0100	0.2253	0.0302	0.3350	0.0081	0.2324
$n = 500$	0.0197	0.2280	0.0048	0.1491	0.0205	0.2320	0.0035	0.1499
$n = 1000$	0.0027	0.1480	-0.0009	0.0896	0.0024	0.1467	-0.0013	0.0903

Table 3: (Design 3) Cross-Sectional Design with $J = 4$ and $p = 3$

	β_1				β_2			
	MEAN	RMSE	MED	MAE	MEAN	RMSE	MED	MAE
$n = 250$	-0.0239	0.5156	-0.0421	0.3790	-0.0116	0.5025	-0.0246	0.3544
$n = 500$	0.0130	0.4239	-0.0033	0.2749	0.0043	0.4219	-0.0139	0.2803
$n = 1000$	0.0113	0.3037	-0.0001	0.1815	0.0114	0.3149	0.0010	0.2001

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 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 seems not the case as we increase the size of the choice set. As seen in Table 3, with 5 alternatives, 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.[25]

We then turn to examine the finite sample properties of the MS estimators for panel data multinomial response models. We start from the static panel model and consider the design (Design 4) with a choice set $\{0, 1, 2\}$ and a panel of two time periods.[26] The indirect utility functions for

[24] To conserve space, we only report the results for β_1 and β_2 in Design 2.

[25] It is not too surprising that our localized estimator performs relatively better in Design 2 than in Design 3. To implement the proposed estimator, we need to match $(J-2) \times p$ binary regressors, which reduces the "effective" sample size. This number for Design 2 is 5, while for Design 3, it is 9. The results presented in Appendix C show that this curse of dimensionality may be alleviated when more of the regressors are continuous.

[26] Intuitively, we would expect that longer panels would improve the finite sample performance of our panel estimators.

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 $\beta_1 = \beta_2 = 1$ are time-invariant, $(x^{(1)}_{i1t})_{i=1,\ldots,n;t=1,2}$ are independent $N(0,1)$ random variables, $(x^{(l)}_{i1t})_{i=1,\ldots,n;t=1,2;l=2,3}$ and $(x^{(l)}_{i2t})_{i=1,\ldots,n;t=1,2;l=1,2,3}$ are independent Bernoulli random variables with parameter 0.5, and

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

The fixed effects are generated as $\alpha_{i1} = (x^{(1)}_{i11} + x^{(1)}_{i12})/4$ and $\alpha_{i2} = (x^{(1)}_{i21} + x^{(1)}_{i22} - 1)/4$. To implement our MS estimator, we use the Epanechnikov kernel and bandwidth $h_n = 6 \cdot (n \log n)^{-1/4}$. In Tables 4 and 5, we report respectively the results for this static panel design using one-step and two-step MS estimators.[27]

[27] That is, we implement our MS method proposed in Section 3 to get $(\hat{\beta}_1, \hat{\beta}_2)$ in the first step estimation, calculate the index $\hat{u}_{ijt} = x^{(1)}_{ijt} + \hat{\beta}_1 x^{(2)}_{ijt} + \hat{\beta}_2 x^{(3)}_{ijt}$, and then run a second step MS estimation by matching on $\hat{u}_{ijs} = \hat{u}_{ijt}$ for $s \neq t$.

Table 4: (Design 4) Static Panel Design with $J = 2$, $p = 3$, and $t \in \{1, 2\}$

	β_1				β_2			
	MEAN	RMSE	MED	MAE	MEAN	RMSE	MED	MAE
$n = 500$	0.0019	0.4974	-0.0065	0.3711	0.0008	0.4897	-0.0053	0.3641
$n = 1000$	0.0244	0.4594	0.0050	0.3067	0.0175	0.4643	0.0004	0.3346
$n = 2000$	0.0199	0.4188	0.0125	0.2720	0.0356	0.4116	0.0159	0.2766
$n = 5000$	0.0125	0.3463	-0.0033	0.2220	0.0160	0.3228	-0.0017	0.2245
$n = 10000$	0.0087	0.2768	0.0000	0.1768	0.0026	0.2664	-0.0032	0.1760

Table 5: (Design 4, Two-step) Static Panel Design with $J = 2$, $p = 3$, and $t \in \{1, 2\}$

	β_1				β_2			
	MEAN	RMSE	MED	MAE	MEAN	RMSE	MED	MAE
$n = 500$	0.0271	0.4601	0.0102	0.3025	0.0475	0.4790	0.0214	0.3364
$n = 1000$	0.0377	0.4268	0.0115	0.2703	0.0194	0.4285	-0.0009	0.2905
$n = 2000$	0.0433	0.3997	0.0166	0.2592	0.0536	0.3892	0.0126	0.2630
$n = 5000$	0.0185	0.3352	-0.0048	0.1954	0.0266	0.3157	0.0023	0.2103
$n = 10000$	0.0120	0.2722	0.0036	0.1738	0.0098	0.2625	-0.0017	0.1720

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

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

where $(\beta, \gamma) = (1, 0.5)$, $y_{i1t-1} = \mathbf{1}[y^*_{i1t-1} > \max\{y^*_{i0t-1}, y^*_{i2t-1}\}]$, $(x^{(1)}_{i1t})_{i=1,\ldots,n;t=0,1,2,3}$ are independent $N(0, 1)$ random variables, $(x^{(2)}_{i1t})_{i=1,\ldots,n;t=0,1,2,3}$ and $(x^{(l)}_{i2t})_{i=1,\ldots,n;t=0,1,2,3;l=1,2}$ are independent Bernoulli random variables with parameter 0.5, $\alpha_{i1} = (x^{(1)}_{i10} + x^{(1)}_{i11} + x^{(1)}_{i12} + x^{(1)}_{i13})/8$ and $\alpha_{i2} = (x^{(1)}_{i20} + x^{(1)}_{i21} + x^{(1)}_{i22} + x^{(1)}_{i23} - 2)/8$, and $(\epsilon_{i1t}, \epsilon_{i2t})_{i=1,\ldots,n;t=0,1,2,3}$ are independent random vectors

drawn from

$$\text{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right).$$

We use the same kernel and bandwidth as the static design above. One-step and two-step estimation results for this design are summarized in Tables 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 sample sizes that are larger than 2000 before the estimator performs adequately well.[28] For the semiparametric 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) Dynamic Panel Design with $J = 2$, $p = 3$, and $t \in \{0, 1, 2, 3\}$

	β				γ			
	MEAN	RMSE	MED	MAE	MEAN	RMSE	MED	MAE
$n = 500$	-0.0270	0.4753	0.0026	0.3292	-0.0412	0.3025	-0.0473	0.2608
$n = 1000$	0.0010	0.4720	0.0020	0.3174	-0.0339	0.2984	-0.0536	0.2542
$n = 2000$	-0.0016	0.4677	0.0121	0.3053	-0.0242	0.2964	-0.0403	0.2512
$n = 5000$	-0.0234	0.4185	-0.0003	0.2225	-0.0214	0.2927	-0.0291	0.2577
$n = 10000$	-0.0107	0.3993	0.0002	0.2143	-0.0104	0.2765	-0.0221	0.2233

Table 7: (Design 5, Two-step) Dynamic Panel Design with $J = 2$, $p = 3$, and $t \in \{0, 1, 2, 3\}$

	β				γ			
	MEAN	RMSE	MED	MAE	MEAN	RMSE	MED	MAE
$n = 500$	-0.0021	0.4205	-0.0051	0.2282	-0.0365	0.3002	-0.0455	0.2631
$n = 1000$	0.0203	0.4158	0.0079	0.2198	0.0055	0.2897	0.0196	0.2502
$n = 2000$	0.0139	0.4051	0.0031	0.2129	-0.0150	0.2795	-0.0420	0.2498
$n = 5000$	0.0456	0.3980	0.0062	0.2108	-0.0211	0.2619	-0.0270	0.2389
$n = 10000$	0.0124	0.3807	-0.0014	0.1999	-0.0111	0.2582	-0.0187	0.2256

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

[28] This result is not that surprising. As all but one of the regressors are binary, our methods use "almost perfectly" matched samples, which only make up a small fraction of the total sample. See Appendix C for simulation results for designs with all continuous regressors.

existence of a continuous regressor, we let all regressors in Design 5 be independent Bernoulli random variables with parameter 0.5. It is easy to show that the Logit estimator in this case would be root-n consistent 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 size of 10000. The inconsistency is to be expected as the Logit estimator is based on i.i.d. type I extreme value errors.

Table 8: (Design 5, Logit) Dynamic Panel Design with $J = 2$, $p = 3$, and $t \in \{0, 1, 2, 3\}$

	β				γ			
	MEAN	RMSE	MED	MAE	MEAN	RMSE	MED	MAE
$n = 500$	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 present an empirical illustration of our proposed estimators by applying it to the often used optical-scanner panel data set on purchases of saltine crackers in the Rome (Georgia) market, 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, the actual price of the purchased brand and shelf prices of other brands, and whether there was a display 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) and 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 the other one 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 row of "display" correspond to fractions of purchase occasions on which each brand is on display. The third explanatory variable we will use is the "price" that corresponds to the price of a brand.[29] 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.[30]

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 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 and panel data (static and dynamic) estimators.

Specifically, here we apply our proposed estimators to the multinomial response model with four alternatives (brands) and three regressors. For each of these estimators, we also implement their corresponding (conditional) likelihood counterparts for comparison. Existing work such as Jain, Vilcassim, and Chintagunta (1994) and Paap and Frances (2000) used this data to estimate multinomial response models with random coefficients. Our semiparametric approach would also be used to examine how sensitive their results and conclusions are to the parametric assumptions they imposed, either in the way of multinomial Logit/Probit specification or modeling unobserved heterogeneity with random coefficients. This can be done by comparing the estimates of the regression coefficients obtained using our methods to those obtained by theirs.

Note that the data set is an unbalanced panel data with $n = 136$ households and the number of purchases varying across households i ($\equiv T_i$, $14 \leq T_i \leq 77$). In what follows, we write $\mathcal{J} = \{1 = \text{Nabisco}, 2 = \text{Sunshine}, 3 = \text{Keebler}, 4 = \text{Private}\}$ for the choice set. For each household i, brand j, and purchase t, we use $x_{ijt}^{(1)}$, $x_{ijt}^{(2)}$, and $x_{ijt}^{(3)}$ to denote the three explanatory variables: the logarithm of "price", "display", and "feature", respectively.

[29] The unit of price in the raw data is cents. Here we convert it to dollars.
[30] We abstract here from issues of endogeneity of price for illustrative purposes.

For the cross-sectional estimator, we model the indirect utility of household i for brand j in the t-th purchase as

$$y^*_{ijt} = -x^{(1)}_{ijt} + \beta_1 x^{(2)}_{ijt} + \beta_2 x^{(3)}_{ijt} + \alpha_j - \epsilon_{ijt},\ j \in \mathcal{J}, t = 1, ..., T_i,$$ [31]

where the coefficient on $x^{(1)}_{ijt}$ is normalized to be -1, α_j is an alternative-specific intercept, and (β_1, β_2) are regression coefficients to be estimated. ϵ_{ijt} is the unobserved scalar disturbance term. The observed choice is defined as $y_{ijt} = \mathbf{1}[y^*_{ijt} > y^*_{ikt}, \forall k \in \mathcal{J}\backslash\{j\}]$. Besides, throughout this section, we denote $x_{ijt} = (x^{(1)}_{ijt}, x^{(2)}_{ijt}, x^{(3)}_{ijt})'$, $\tilde{x}_{ijt} = (x^{(2)}_{ijt}, x^{(3)}_{ijt})'$, $x_{it} = (x'_{i1t}, x'_{i2t}, x'_{i3t}, x'_{i4t})'$, $x_i = (x'_{i1}, ..., x'_{iT_i})'$, $y_{it} = (y_{i1t}, y_{i2t}, y_{i3t}, y_{i4t})'$, and $y_i = (y'_{i1}, ..., y'_{iT_i})'$.

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 of the following form with respect to (b_1, b_2).

$$\sum_i \sum_{s=1}^{T_i-1} \sum_{t>s} K_{h_n}(x^{(1)}_{i(-1)s} - x^{(1)}_{i(-1)t}) \mathbf{1}[\tilde{x}_{i(-1)s} = \tilde{x}_{i(-1)t}](y_{i1s} - y_{i1t}) \cdot sgn((x_{i1s} - x_{i1t})'b)$$
$$+ \sum_{i \neq m} \sum_{s=1}^{T_i} \sum_{t=1}^{T_m} K_{h_n}(x^{(1)}_{i(-1)s} - x^{(1)}_{m(-1)t}) \mathbf{1}[\tilde{x}_{i(-1)s} = \tilde{x}_{m(-1)t}](y_{i1s} - y_{m1t}) \cdot sgn((x_{i1s} - x_{m1t})'b), \quad (5.1)$$

where $b = (-1, b_1, b_2)'$ and $x^{(1)}_{i(-1)t}$ ($\tilde{x}_{i(-1)t}$) denotes the vector collecting $x^{(1)}_{ijt}$ ($\tilde{x}_{ijt}$) for all $j \in \mathcal{J}\backslash\{1\}$. To compute (5.1), we use the fourth-order Gaussian kernel function and bandwidth $h_n = c \cdot \hat{\sigma} n^{-1/7}$, where $\hat{\sigma}$ is the standard deviation of the matching variable and c is a constant. These choices consider the scale of the data and satisfy Assumptions CS6 and CS8–CS9. To test the sensitivity of our method to the choice of bandwidths, we experiment with several values of c, ranging from 0.5 to 2. Results were not sensitive to this choice, so we only report results for $c = 1$.

To attain confidence intervals (CI) we employ the nonparametric bootstrap for clustered data. Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$ denote the MRC estimate of $\beta = (\beta_1, \beta_2)'$. The 95% confidence intervals for β_1 and β_2 are constructed by the following algorithm:

1. Draw $(y^{*\prime}_i, x^{*\prime}_i)'$, $i = 1, ..., n$, independently with replacement from the original sample.[32]
2. Estimate β from objective function (5.1), using the bootstrap sample obtained in step 1.
3. Repeat steps 1 and 2 for $B = 200$ times to get a series of estimates of β, $\{\hat{\beta}^*_{(b)}\}_{b=1}^B$.

[31] Note that one can always obtain an expression as in (2.1) by subtracting the indirect utility for one (base) alternative from the other indirect utilities. Our estimator is numerically invariant to the choice of the base alternative.

[32] Note that here we estimate a pooled model using short panel data. To account for heteroskedasticity across "clusters" of observations (i.e., household i), our bootstrap resamples the cluster i rather than both i and t.

4. Let $Q_{\hat{\beta}^*}(\tau)$ denote the τ-th quantile of $\{\hat{\beta}^*_{(b)}\}_{b=1}^B$, $0 \leq \tau \leq 1$. The 95% confidence interval for β can be constructed as $[\hat{\beta} - (Q_{\hat{\beta}^*}(0.975) - \hat{\beta}),\ \hat{\beta} - (Q_{\hat{\beta}^*}(0.025) - \hat{\beta})]$. Alternatively, one can use the normal approximation by computing $[\hat{\beta}_l - 1.96 \times \hat{\text{se}}(\hat{\beta}_l), \hat{\beta}_l + 1.96 \times \hat{\text{se}}(\hat{\beta}_l)]$ for $l = 1, 2$, where

$$\hat{\text{se}}(\hat{\beta}_l) = \sqrt{\frac{1}{B-1}\sum_b^B (\hat{\beta}^*_{l(b)} - \bar{\hat{\beta}}^*_{l(b)})^2}$$

with $\bar{\hat{\beta}}^*_{l(b)} = B^{-1}\sum_b^B \hat{\beta}^*_{l(b)}$.

Point estimates and confidence intervals for each of the two coefficients on the binary regressors are reported in Table 10. For the semiparametric estimator, we place the 95% confidence regions on top of the normal approximation results. For comparison purposes, the table also reports results from estimators for two (pooled) parametric models, multinomial Logit and multinomial Probit.[33] Recall that the two parametric methods and our semiparametric estimator impose different scale normalization. To facilitate the comparison, we report the ratios of the coefficients of the two binary regressors to the absolute value of the coefficient on $x^{(1)}_{ijt}$.

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	0.0166	(-0.0246, 0.0352)	0.1192	(0.0765, 0.2266)
		(-0.0227, 0.0559)		(0.0350, 0.2034)
Multinomial Logit	0.0330	(-0.0513, 0.1174)	0.1573	(-0.0029, 0.3175)
Multinomial Probit	0.0155	(-0.0437, 0.0747)	0.1108	(0.0369, 0.1847)

As we can see, these results are not quite different. For the parametric estimators, multinomial Probit relative coefficients for $x^{(2)}_{ijt}$ and $x^{(3)}_{ijt}$ are 0.0155 and 0.1108, respectively, though only the latter is significantly different from 0 at the 95% level. For the multinomial Logit estimator, the corresponding coefficient ratio estimates are of larger magnitude, i.e., $(0.0330, 0.1573)$. However, in contrast to the Probit results, none of them are significantly different from 0 at the 95% level. Our semiparametric estimator gives very similar estimation and inference results as Probit.

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

$$y^*_{ijt} = -x^{(1)}_{ijt} + \beta_1 x^{(2)}_{ijt} + \beta_2 x^{(3)}_{ijt} + \alpha_{ij} - \epsilon_{ijt},\ j \in \mathcal{J}, t = 1, ..., T_i,$$

where α_{ij} collects the individual and alternative specific effects. Note that here α_{ij} can be arbitrarily

[33] We implement these estimations in Stata and compute the standard errors using Stata's cluster-robust option.

correlated with x_i. The objective function of our semiparametric estimator is of the form

$$\sum_i \sum_{s=1}^{T_i-1} \sum_{t>s} K_{h_n}(x^{(1)}_{i(-1)s} - x^{(1)}_{i(-1)t}) \mathbf{1}[\tilde{x}_{i(-1)s} = \tilde{x}_{i(-1)t}](y_{i1s} - y_{i1t}) \cdot sgn((x_{i1s} - x_{i1t})'b). \tag{5.2}$$

To compute (5.2), we choose the Gaussian kernel function and $h_n = 3\hat{\sigma} n^{-1/6} / \sqrt[3]{\log n}$. In addition to our estimator, we also implement the conditional likelihood estimator proposed in Chamberlain (1980) for comparison. The conditional likelihood method is consistent for the Logit specification, but it may not be consistent for general specifications.

As presented in Table 11, the estimation results for our semiparametric method are$(\hat{\beta}_1, \hat{\beta}_2) = (0.0639, 0.0822)$. 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 coefficients on $x^{(2)}_{ijt}$ and $x^{(3)}_{ijt}$ are comparable. This is in complete contrast to cross-sectional models where the coefficient on $x^{(2)}_{ijt}$ is not statistically different from 0 and the coefficient on $x^{(3)}_{ijt}$ is significantly positive. The semiparametric estimates are also strikingly different from the conditional Logit estimates,[34] indicating that the Logit model may be misspecified. However, for the panel data model, we only report point estimates but not confidence intervals. This is because the limiting distribution of our panel estimator is nonstandard, and thus it is unlikely that the standard bootstrap can provide valid inference results in this setting.

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

	β_1	β_2
Semiparametric	0.0639	0.0822
Conditional Logit	0.0865	0.2271

The last task for this section is to examine the state dependence in the panel data model. Particularly, we consider the following model modified from the static panel setting: For each i and $t \in \{2, ..., T_i\}$,

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

i.e., we include $y_{i1(t-1)}$ in the indirect utility of alternative 1 as an additional regressor. To estimate $(\beta_1, \beta_2, \gamma)$, we work with the following objective functions, generalizing the semiparametric and conditional Logit estimators proposed in Section 3.2 respectively to models with a longer (unbalanced)

[34] Similar to the cross-sectional model, all parametric estimates reported in Tables 11 and 12 (see below) are the ratios of the coefficients on the other regressors to the absolute value of the coefficient on $x^{(1)}_{ijt}$.

panel. Let $v_{it}^{(1)} \equiv (x_{i1t}^{(1)}, x_{i(-1)t}^{(1)\prime})'$ and $\tilde{v}_{it} \equiv (\tilde{x}_{i1t}', \tilde{x}_{i2t}', \tilde{x}_{i3t}', \tilde{x}_{i4t}')'$. They are

$$\begin{aligned}
&\sum_i \sum_{t=2}^{T_i-2} K_{h_n}(x_{i(-1)t}^{(1)} - x_{i(-1)t+1}^{(1)}, v_{it+1}^{(1)} - v_{it+2}^{(1)}) \mathbf{1}[\tilde{x}_{i(-1)t} = \tilde{x}_{i(-1)t+1}, \tilde{v}_{it+1} = \tilde{v}_{it+2}] \\
&\quad \times (y_{i1t} - y_{i1t+1}) \cdot sgn((x_{i1t} - x_{i1t+1})'b + r(y_{i1t-1} - y_{i1t+2})) \\
&+ \sum_i \sum_{s=2}^{T_i-3} \sum_{t=s+2}^{T_i-1} K_{h_n}(x_{i(-1)s}^{(1)} - x_{i(-1)t}^{(1)}, v_{is+1}^{(1)} - v_{it+1}^{(1)}) \mathbf{1}[\tilde{x}_{i(-1)s} = \tilde{x}_{i(-1)t}, \tilde{v}_{is+1} = \tilde{v}_{it+1}] \\
&\quad \times \mathbf{1}[y_{i1s+1} = y_{i1t+1}](y_{i1s} - y_{i1t}) \cdot sgn((x_{i1s} - x_{i1t})'b + r(y_{i1s-1} - y_{i1t-1}))
\end{aligned}$$

and

$$\begin{aligned}
&\sum_i \sum_{2 \le s < t \le T_i - 1} K_{h_n}(x_{i(-1)s}^{(1)} - x_{i(-1)t}^{(1)}, v_{is+1}^{(1)} - v_{it+1}^{(1)}) \mathbf{1}[\tilde{x}_{i(-1)s} = \tilde{x}_{i(-1)t}, \tilde{v}_{is+1} = \tilde{v}_{it+1}] \mathbf{1}[y_{i1s} \neq y_{i1t}] \\
&\times \log\left(\frac{\exp((x_{i1s} - x_{i1t})'b + r(y_{i1s-1} - y_{i1t+1}) + r(y_{i1s+1} - y_{i1t-1})\mathbf{1}[t-s>1])^{y_{i1s}}}{1 + \exp((x_{i1s} - x_{i1t})'b + r(y_{i1s-1} - y_{i1t+1}) + r(y_{i1s+1} - y_{i1t-1})\mathbf{1}[t-s>1])} \right).
\end{aligned}$$

To implement these two estimators,[35] we consider the Gaussian kernel function and $h_n = 5\hat{\sigma}n^{-3/29}$.

As reported in Table 12 below, the estimation results are $(\hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}) = (0.1358, 0.2460, 0.2512)$ for the semiparametric estimator, and $(\hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}) = (-0.1481, 0.2815, 0.2353)$ for the conditional Logit estimator. For the semiparametric 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 the state dependence may 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. The conditional Logit model is probably misspecified for this data set as the estimated coefficient on $x_{ijt}^{(2)}$ is negative with relatively large magnitude, not making economic sense.

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

	β_1	β_2	γ
Semiparametric	0.1358	0.2460	0.2512
Conditional Logit	-0.1481	0.2815	0.2353

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

[35] For the conditional Logit estimator, the first component of b is a free parameter to be estimated.

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 panel data estimator of a model with alternative 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 both strict exogenous and lagged dependent variables.

The work here leaves many open areas for future research. For example, as pointed out, in both panel data settings, the proposed procedure suffers from a curse of dimensionality in the number of choices. It is thus an open question if our proposed approach results 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.

Abrevaya, J. and J. Huang (2005): "On the bootstrap of the maximum score estimator," *Econometrica*, 73, 1175–1204.

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.

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.

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics: methods and applications*, Cambridge university press.

Cattaneo, M. D., M. Jansson, and K. Nagasawa (2020): "Bootstrap-Based Inference for Cube Root Asymptotics," *Econometrica*, 88, 2203–2219.

Cavanagh, C. and R. P. Sherman (1998): "Rank estimators for monotonic index models," *Journal of Econometrics*, 84, 351–382.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *The Review of Economic Studies*, 47, 225–238.

——— (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.

Charlier, E. (1997): "Limited dependent variable models for panel data," Tech. rep., Tilburg University, School of Economics and Management.

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. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and quantile effects in nonseparable panel models," *Econometrica*, 81, 535–580.

Delgado, M., J. Rodríguez-Poo, and M. Wolf (2001): "Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator," *Economics Letters*, 73, 241–250.

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.

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

Hong, H. and J. Li (2020): "The numerical bootstrap," *The Annals of Statistics*, 48, 397–412.

Hong, H., A. Mahajan, and D. Nekipelov (2015): "Extremum estimation and numerical derivatives," *Journal of Econometrics*, 188, 250–263.

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

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

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.

Jin, Z., Z. Ying, and L. Wei (2001): "A simple resampling method by perturbing the minimand," *Biometrika*, 88, 381–390.

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.

Kyriazidou, E. (1997): "Estimation of a panel data sample selection model," *Econometrica: Journal of the Econometric Society*, 1335–1364.

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

Lee, S. M. S. and M. C. Pun (2006): "On m out of n bootstrapping for nonstandard m-estimation with nuisance parameters," *Journal of American Statistical Association*, 101, 1185–1197.

Magnac, T. (1997): "State Dependence and Heterogeity in youth Employment Histories," Tech. rep., uCL working paper.

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

——— (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27(3), 313–33.

——— (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.

Ouyang, F. and T. T. Yang (2020a): "Semiparametric Discrete Choice Models for Bundles," Tech. rep., University of Queensland, School of Economics.

——— (2020b): "Semiparametric Estimation of Dynamic Binary Choice Panel Data Models," Tech. rep., University of Queensland, School of Economics.

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.

Rasch, G. (1960): "Probabilistic models for some intelligence and achievement tests," *Copenhagen: Danish Institute for Educational Research.*

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.

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

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

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

Subbotin, V. (2007): "Asymptotic and bootstrap properties of rank regressions," *Available at SSRN 1028548*.