Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity

Manu Navjeevan

May 29, 2020

1 Introduction

Authors are Xiaoxia Shi (Wisconsin), Matthew Shum (CalTech), Wei Song (Xiamen University)

Paper proposes new semi-parametric identification and estimation appraoch to multinomial choice models in a panel data setting with individual fixed effects. Approach based on *cyclic monotnoicity*, which helps derive identifying inequalities without requiring shape restrictions for distribution of underlying random utility shock.

Consider a panel multinomial choice problem. Agent i chooses from $K+1$ options (labeled $k = 0, \ldots, K$). Choosing option k in period t gives the agent indirect utility

$$\beta' X_{it}^k + A_i^k + \epsilon_{it}^k \tag{1.1}$$

where X_{it}^k is a d_x-dimensional vector of observable covariates with support $\mathcal{X}$. A_i^k is a agent specific fixed effect and ϵ_{it}^k are unobservable utility shocks with unspecified distribution. Agent shookses option that gives highest utility

$$Y_{it}^k = \mathbb{1}\{\beta' x_{it}^k \geq \beta' X_{it}^{k'} + A_i^{k'} + \epsilon_{it}^{k'}\} \tag{1.2}$$

Assume that the data is i.i.d across $i \in \mathcal{I}$. Normalize $\|\beta\| = 1$, $X_{it}^0 = \mathbf{0}_{d_x}$, and $A_i^0 = \epsilon_{it}^0 = 0$. Do not impose any location normalization on ϵ_{it}^k or A_i^k so WLOG X_{it}^k does not contain a constant.

Expoit a notion of *cyclic monotonicity*, which is an appropriate generalization of monotonicity to the identification and estimation of semi-parametric multinomial choice models. Cyclic monotonicity creats bounds which can then be used to identify parameters.

2 Preliminaries

Definition 1. (Cyclic Monotonicity) Consider a function $f : \mathcal{U} \to \mathbb{R}^k$ where $\mathcal{U} \subset \mathbb{R}^k$ and a length M cycle of points in $\mathbb{R}^k : u_1, u_2, \ldots, u_m, u_1$. The function f is cyclic monotone if

$$\sum_{m=1}^{M} (u_m - u_{m+1})' f(u_m) \geq 0 \tag{2.1}$$

where $u_{M+1} = u_1$. The function f is cyclic monotone on $\mathcal{U}$ if it is cyclic monotone with respect to all possible cycles of all lengths on it's domain.[1]

Cyclic monotonicity is defined for mappings from $\mathbb{R}^K \to \mathbb{R}^K$, which generalizes the usual monotnicity for real-valued functions.

[1] This seems just a little bit weaker than saying that f is monotonic. If f takes values in $\mathbb{R}$ and U is compact, these seem equivalent. I think the weakening in higher dimensions is that it is close to requiring monotonicity for any projection mapping $\pi(f)(\cdot)$

Proposition 1. *(Cyclic Monotonicity and Convexity) Consider a differentiable function $F : \mathcal{U} \to \mathbb{R}$ for an open convex set $\mathcal{U} \subset \mathbb{R}^K$. If F is convex on $\mathcal{U}$, then the gradient of F, denoted $\nabla F(u) = \partial F(u)/\partial u$, is cyclic monotone on $\mathcal{U}$.*

Proof should follow quickly from the fact that $\nabla^2 F(u)$ is positive semidefinite.

To connect the above defiinitons to the multinomial choice model, start with a generic random utility model for multinomial choices without specifying the random utility function or the data structure in detail. Suppose an agent is choosing from $K+1$ choices $\{0, 1, \ldots, K\}$. Utility from choice k is partitioned into additive parts $U^k + \epsilon^k$, where U^k denotes the systematic component of the latent utility where ϵ^k denotes random shocks, idiosyncratic across agents and choice occasions. She chooses choice k^* if

$$U^{k^*} + \epsilon^{k^*} \geq \max_{k=0,\ldots,K} U^k + \epsilon^k$$

Let $Y^k = 1$ if choice k is taken and 0 otherwise. Normalize utility of outside option to 0. Further let u^k denote a generic realization of U^k. Also let $U = (U^1, \ldots, U^K)'$, $\mathbf{u} = (u^1, \ldots, u^K)'$ and $\epsilon = (\epsilon^1, \ldots, \epsilon^K)'$. Introduce the "social surplus function" from McFadden (1978, 1981). which is expected utility obtained from the choice problem

$$\mathcal{W}(\mathbf{u}) = \mathbb{E}\left\{ \max_{k=0,1,\ldots,K} \left[U^k + \epsilon^k\right] |\mathbf{U} = \mathbf{u} \right\} \tag{2.2}$$

Following lemme shows that this function is convex and differentiable, gradient corresponds to the choice probability function, and that the choice probability function is cyclic monotone[2].

Lemma 1. *(Gradient) Suppose that $\mathbb{U}$ is independent of ϵ and that the distribution of ϵ is absolutely continuous with respect to the Lebesgue measure. Then*

1. $\mathcal{W}(\cdot)$ *is convex on* $\mathbb{R}^K$
2. $\mathcal{W}(\cdot)$ *is differentiable on* $\mathbb{R}^K$,
3. $\mathbf{p}(\mathbf{u}) = \nabla \mathcal{W}(\mathbf{u})$. *where* $\mathbf{p}(\mathbf{u}) = \mathbb{E}[Y|\mathbf{U} = \mathbf{u}]$ *and* $\mathbf{Y} = (Y^1, \ldots, Y^K)'$
4. $\mathbf{p}(\mathbf{u})$ *is cyclic monotone on* $\mathbb{R}^K$

Cyclic monotonicity of the choice probability can be used to form identifying restrictions for the structural parameters in a variety of setttings. Paper focuses on linear panel data model with fixed effects, described in section one above.

3 Panel Data Multinomial Choice Models with Fixed Effects

Focus on a short panel data setting with only two time periods. Extension to multiple time periods is given in section 5. Let $\mathbf{U}, \epsilon and \mathbf{Y}$ be indexed by both i and t (individual and time period). Let there be an observable d_x-dimensional covariate X_{it}^k for each choice k, and let U_{it}^k be a linear index of X_{it}^k plus an unobservable individual effect A_i^k, where β is a d_x-dimensional unknown parameter. Let $\mathbf{X}_{it} = (X_{it}^1, \ldots, X_{it}^K)$ and $\mathbf{A}_i = (A_i^1, \ldots, A_i^K)'$. $\mathbf{X}_{it}$ is a $d_x \times K$ matrix. In short panels, the challenge is the identification of β while allowing correlation between the covariates and the individual effects. Tackle this problem using the cyclic monotonicity of the choice probability.

3.1 Identifying Inequalities

Derive inequalities under the following assumption

Assumption 1. 1. ϵ_{i1} and $\epsilon_{\mathbf{i2}}$ are identically distributed conditional on $\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}$:

$$(\epsilon_{i1} \sim \epsilon_{i2})|(\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2})$$

[2] This clearly follows from Proposition 1.

2. The conditional distribution of ϵ_{it} given $\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}$ is absolutely continuous with respect to the Lebesgue measure for $t = 1, 2$ everywhere on the support of $\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}$

Part 1 of the above assumption is the multinomial version of the group homogeneity assumption of Manski (1987) and is also imposed by Pakes and Porter (2013). It allows us to form identification inequalities based on the comparasion of choices made by the same individual over different time period, and, by doing so, eliminate the fixed effect. This assumption rules out dynamic panel models where X_{it}^k may include lagged values of Y_{it}^k, but it allows $\epsilon_{\mathbf{it}}$ to be correlated with the covariates, and allows arbitrary dependence between ϵ_{it} and the fixed effects. The second part of the above assumption imposes no restriction on the dependence amongst the errors. Errors across choices in a given period can be arbitrarily dependent, and the errors across time periods, altough assumed to have identical marginal disributions, can have arbitrary dependence.

Now to show identification (partial), begin by letting η be a K-dimensional vector with k-th element η^k and define

$$\mathbf{p}(\eta, \mathbf{x_1}, \mathbf{x_2}, \mathbf{a}) := \left(\mathbb{P}\left[\epsilon_{i1}^k = \eta^k \geq \epsilon_{i1}^{k'} + \eta^{k'}\ \forall k' \middle| \mathbf{X_{i1}} = \mathbf{x_1}, \mathbf{X_{i2}} = \mathbf{x_2}, \mathbf{A_i} = \mathbf{a}\right]\right)_{k=1,\ldots,K} \tag{3.1}$$

Assumption 1.1 implies that

$$\mathbf{p}(\eta, \mathbf{x_1}, \mathbf{x_2}, \mathbf{a}) := \left(\mathbb{P}\left[\epsilon_{i2}^k = \eta^k \geq \epsilon_{i2}^{k'} + \eta^{k'}\ \forall k' \middle| \mathbf{X_{i1}} = \mathbf{x_1}, \mathbf{X_{i2}} = \mathbf{x_2}, \mathbf{A_i} = \mathbf{a}\right]\right)_{k=1,\ldots,K} \tag{3.2}$$

whereas Assumption 1.2 along with Lemma 1 imply that $\mathbf{p}(\eta, \mathbf{x_1}, \mathbf{x_2}, \mathbf{a})$ is cyclic monotone in η for all possible values of $\mathbf{x_1}, \mathbf{x_2}, \mathbf{a}$

Using cyclic monotonicity for length 2 cycles we obtain, for any $\eta_\mathbf{1}, \eta_\mathbf{2}$ and $\mathbf{x_1}, \mathbf{x_2}, \mathbf{a}$ we have

$$(\eta_1 - \eta_2)' \left[\mathbf{p}(\eta_\mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{a}) - \mathbf{p}(\eta_\mathbf{2}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{a})\right] \geq 0 \tag{3.3}$$

Now let $\eta_1 = X_{i1}'\beta + \mathbf{A}_i$ and $\eta_2' = X_{i2}'\beta + A_i$. By the definition of $\mathbf{p}(\cdot)$ we have

$$\mathbf{p}\left(X_{it}'\beta + A_i, X_{i1}, X_{i2}, A_i\right) = \mathbb{E}\left[\mathbf{Y}_{it} \middle| \mathbf{X_{i1}}, \mathbf{X_{i2}}, \mathbf{A_i}\right] \tag{3.4}$$

Combining the above we have that

$$\left(\mathbb{E}\left[\mathbf{Y}_{i1}' \middle| \mathbf{X_{i1}}, \mathbf{X_{i2}}, \mathbf{A_i}\right] - \mathbb{E}\left[\mathbf{Y}_{i2}' \middle| \mathbf{X_{i1}}, \mathbf{X_{i2}}, \mathbf{A_i}\right]\right)\left(\mathbf{X}_{i1}'\beta - \mathbf{X}_{i2}'\beta\right) \geq 0 \ \text{ everywhere} \tag{3.5}$$

Fixed effect within the second paranthetical term on LHS defferences out. So can take conditional expectation given the X values to obtain

$$\left(\mathbb{E}\left[\mathbf{Y}_{i1}' \middle| \mathbf{X_{i1}}, \mathbf{X_{i2}}\right] - \mathbb{E}\left[\mathbf{Y}_{i2}' \middle| \mathbf{X_{i1}}, \mathbf{X_{i2}}\right]\right)\left(\mathbf{X}_{i1}'\beta - \mathbf{X}_{i2}'\beta\right) \geq 0 \ \text{ everywhere} \tag{3.6}$$

Last equation only involves identifies/observed quantities and the unknown parameter β. Under $K = 1$ (binary choice), the inequalities reduce to the rank correlation result in Maski (1987, Lemma 1).

Extension in Section 5 discusses how longer cycles can be used when more time periods are available in the data set. Next consider point identification

3.2 Point Identification of Model Parameters

Rewrite (3.6) as

$$\mathbb{E}[\Delta Y_i'|X_{i1},X_{i2}]\Delta X_i'\beta \geq 0 \tag{3.7}$$

Define $g = (\Delta X_i \mathbb{E}[\Delta Y_i|X_{i1},X_{i2}])$. For identification, want to place restrictions on the support of the vector g which we define as

$$\mathcal{G} = \operatorname{supp}(g) = \operatorname{supp}(\Delta X_i \mathbb{E}[\Delta Y_i|X_{i1},X_{i2}]) \tag{3.8}$$

Want to find conditions on model primitives that guarantee that the support of g is rich enough to ensure point identification. First impose some regularity conditions

Assumption 2. 1. The conditional support of $\epsilon_{it}|A_i,X_{i1},X_{i2}$ is $\mathbb{R}^K$ with positive probability everywhere

2. The conditional distribution of $(\epsilon_{it}+A_i)$ given $(X_{i1},X_{i2})=(x_1,x_2)$ is uniformly continuous in (x_1,x_2). That is,

$$\lim_{(x_1,x_2)\to(x_1^0,x_2^0)} \sup_{e,a\in\mathbb{R}^K} \left| F_{\epsilon_{it}+A_i|X_{i1},X_{i2}}(e+a|x_1,x_2) - F_{\epsilon_{it}+A_i|X_{i1},X_{i2}}(e+a|x_1^0,x_2^0)\right| = 0$$

Assumption 2.2 is a suffecient condition for the continuity of the function $\mathbb{E}[\Delta Y_i|X_{i1},X_{i2}]$. This ensures that the violation of the inequality $\mathbb{E}[\Delta Y_i|X_{i1}=x_1,X_{i2}=x_2]\Delta x'b \geq 0$ for a point (x_1,x_2) on the support of (X_{i1},X_{i2}) implies that the inequality $\mathbb{E}[\Delta Y_i'|X_{i1},X_{i2}]\Delta X_i'b \geq 0$ is violated with positive probability[3]

Also need a condition on X. All vectors g are equal to

$$\Delta X_i E[\Delta Y_i|X_{i1},X_{i2}] = \sum_{k=1}^{K} \Delta X_i^k E[\Delta Y_i^k|X_{i1},X_{i2}]$$

In general it is difficult to formulate conditions on RHS of the equation becasue RHS is a weighted sum of ΔX_i^k where the weight is the conditional choice probability, which is not a primitive quantity. Proceed by considering two approaches to reduce RHS to a single term.

1. For a given k, let $\Delta X_i^{-k} = (\Delta X_i^1,\ldots,\Delta X_i^{k-1},\Delta X_i^{k+1},\ldots,\Delta X_i^K)$. Conditional on the event $\Delta X_i^{-k}=0$, we have

$$\Delta X_i \mathbb{E}[\Delta Y_i|X_{i1},X_{i2}] = \Delta X_i^k \mathbb{E}[\Delta Y_i^k|X_{i1},X_{i2}]$$

Because $\mathbb{E}[\Delta Y_i^k|X_{i1},X_{i2}]$ is a scalar random variable,

$$\operatorname{supp}(\Delta X_i^K \mathbb{E}[\Delta Y_i^k|X_{i1},X_{i2}]) = \operatorname{supp}(\Delta X_i^K \operatorname{sign}\mathbb{E}[\Delta Y_i^k|X_{i1},X_{i2}])$$

Assumption 3.2(a) ensures that $\mathbb{P}(\mathbb{E}[\Delta Y_i^K|X_{i1},X_{i2}]=0|\Delta X_i^{-k}=0)=0$, which implies that

$$\operatorname{sign}(\mathbb{E}[\Delta Y_i^k|X_{i1},X_{i2}]) \in \{-1,1\} \text{ with pr. } 1$$

So, it is suffecient to assume a rich support for Δ_i^k and $-\Delta X_i^k$ conditional on $\Delta X_i^k=0$. We are thus motivated to define support for ΔX_i^k and $-\Delta X_i^k$ conditonal on $\Delta X_i^k=0$. Gives motivation to define

$$G_I = \bigcup_k \operatorname{supp}(\pm\Delta X_i^k|\Delta X_i^{-k}=0) \tag{3.9}$$

[3] The violation of $\mathbb{E}[\Delta Y_i|X_{i1}=x_1,X_{i2}=x_2]\Delta x'b \geq 0$ at a point, along with continuity, implies a violation of the inequality in a neighborhood around that point. It's not clear to me why the neighborhood have positive measure. It seems to me that

$$f(x) = \begin{cases} 1 & \text{if } x\in[0,1] \\ 100 & \text{if } x=100 \\ 0 & \text{otherwise} \end{cases}$$

would allow for continuity without the neighborhood around $x=100$ having positive measure. Maybe though, the function $\mathbb{E}[\Delta Y_i|X_{i1}=x_1,X_{i2}=x_2]$ would not be defined in a neighborhood around $x=100$ in this case and so we're fine.

where the conditional support of $\pm\Delta X_i^k$ is the union of the conditoinal support of ΔX_i^k and that of $-\Delta X_i^k$.

2. Conditional on the event $\Delta X_i^k = \Delta X_i^1$ for all k (individual $i's$ covariates are identical across all choices and only vary across time periods) we have

$$\Delta X_i \mathbb{E}[\Delta Y_i | X_{i1}, X_{i2}] = \Delta X_i^1 E[-\Delta Y_i^0]$$

where $\Delta Y_i^0 = \sum_{k=1}^K \Delta Y_i^k$. Similar arguments as above show that it is suffecient to assume a rich support for ΔX_i^1, which motivates the definition

$$G_{II} = \bigcup_k \text{supp}(\pm\Delta X_i^k | \Delta X_i^k = \Delta X_i^1, \forall k) \tag{3.10}$$

Identification condition is imposed on the set $G \equiv G_I \cup G_{II}$. Two assumptions on G are considered which differ in the types of covariats they accomodate. Each assumption is sufficient by itself. Consider each case in turn.

Assumption 3. The set G contains an open R^{d_x} ball around the origin.

Roughly, beginning at the origin and moving in any direction, we will reach a point in G. This assumption essentially requires all covariates to be continuous, but allows them to be bounded (usually assume unboundedness to obtain a result).

Assumption 4. For some $j^* \in \{1, 2, \ldots, d_x\}$:

1. $G_{j^*}(g_{-j^*}) = R$ for all g_{-j^*} in a subset G_{-j}^0 of G_{-j^*}
2. $G_{-j^*}^0$ is not contained in a proper linear subspace of $\mathbb{R}^{d_x - 1}$
3. the j^*-th of β, denoted β_{j^*} is nonzero.

Identification result is stated using the following criterion function:

$$Q(b) = \mathbb{E}\left|\min\left(0, \mathbb{E}[\Delta Y_i' | X_{i1}, X_{i2}]\Delta X_i' b\right)\right| \tag{3.11}$$

Which will be returned to in considering estimation.

Theorem 1. *Under Assumptions 1,2, and either 3 or 4, we have $Q(\beta) = 0$ and $Q(b) > 0$ for all $b \neq \beta$ such that $b = \mathbb{R}^{d_x}$ and $\|b\| = 1$*

3.3 Examples

Example 1: $\text{supp}((X_{it}^k)_{t=1,2,k=1,2}) = [0,1]^8$ then $\text{supp}((\Delta_i^k)_{k=1,2}) = [-1,1]^4$. In this case Assumption 3 is clearly satisfied.

Example 2: Suppose that the covariates do not vary across k (for example coordinates are individual specific), that is $X_{it}^k = X_{it}$ for $k = 1, 2$ and $\text{supp}((X_{it})_{t=1,2}) = [0,1]^4$. Thus $G_{11} = \text{supp}(\Delta X_i) = [-1,1]^2$, satisfying Assumption 3.3.

Example 3: Suppose that the covariates take continuous values for alternative 1 and discrete values for alternative 2. For example, $\text{supp}((X_{it}^1)_{t=1,2}) = [0,1]^4$ and $\text{supp}((X_{it}^2)_{t=1,2}) = \{0,1\}^4$ and the joint support is the cartesian product. Then $\text{supp}(\Delta X_i^1 | \delta X_i^2 = 0) = [-1,1]^2$. So Assumption 3 is satisfied .

Example 4: Suppose that the first covariate is a time dummy: $X_{1,it}^k = t$ for all k, t and the second covariate has unbounded support $\text{supp}((X_{2,it}^k)_{t=1,2,k=1,2})) = (c, \infty)^4$ for some $c \in \mathbb{R}$. Then,

$$\text{supp}(\Delta X_i^1 | \Delta X_i^1 = \Delta X_i^2) = \{1\} \times \mathbb{R}$$

so $G \supseteq G_{II} = \{-1, 1\} \times \mathbb{R}$. Let $j^* = 2$ and $G^0_{-2} = \{-1, 1\}$. Then Assumption 4.2 holds and Assumption 4.1 holds becouse $G_2(-1) = G_2(1) = \mathbb{R}$. Assumption 4.3 holds as long as $\beta_2 \neq 0$

3.4 Remarks: Cross Sectional Model

Paper has so far focused on identification of a *panel* multinomial choice model. Breifly remark here that the cuclic monotonicity inequalities can be used for estimation in cross-sectional multinomial choice models, which is natural and can be compared to the large number of existing estimators for these models. In the cross-sectional model, the individual specific fixed effects disappear, leading to the choice model

$$Y_i^k = \mathbb{1}\{\beta' X_i^k + \epsilon_i^k \geq \beta' X_i^{k'} + \epsilon_i^{k'} \forall k'\}$$

so to apply the CM inequalities, the only dimension upon which one can difference is across individuals. Under the assumptions that the vecotr of utility shocks ϵ_i is (i) i.i.d across individuals and (ii) independent of the covariates X, the 2-cycle CM inequaly yields that, for all pairs (i,j)

$$\left(\mathbb{E}[Y_i|X_i] - E[Y_j|X_j]\right)' (X_i - X_j)'\beta \geq 0$$

In particular, for the binary choice case, this reduces to

$$\left(\mathbb{E}[Y_i^1|X_i] - E[Y_j^1|X_j]\right)' (X_i - X_j)'\beta \geq 0$$

which is the estimating equation underlying the maximum score and rank correlation estimators for the binary choice model.

4 Estimation and Consistency

Section proposes a computationall yeasy and consistent estimator based on Theorem 1. Consistency is obtained with $T \to \infty$ with T fixed. In particular focus on $T = 2$ and only discuss longer panels in next section. Consistency should follow from consistency of GMM (and consistency of M-estimators more generally). Specifically obtain a consistent estimator of β, $\hat{\beta} = \bar{\beta}/\|\bar{\beta}\|$ where

$$\bar{\beta} = \arg \min_{b \in \mathbb{R}^{d_x} : \max_j |b_j| = 1} Q_n(b)$$

with

$$Q_n(b) = n^{-1} \sum_{i=1}^{n} \left[\left(b' \Delta X_i\right) \left(\Delta \hat{p}(X_{i1}, X_{i2})\right) \right]_-$$

here $\Delta \hat{p}(X_{i1}, X_{i2}) = \hat{p}_2(X_{i1}, X_{i2}) - \hat{p}(X_{i1}, X_{i2})$ and $\hat{p}_t(x_1, x_2)$ is a consistent[4] estimator for $\mathbb{E}(Y_{it}|X_{i1}, X_{i2})$

Assumption 5. Assume that:

1. $\max_i \|\hat{p}(\cdot) - p(\cdot)\| \to_p 0$ is uniformly consistent and
2. $\max_{t=1,2} \mathbb{E}[\|X_{it}\|] < \infty$

Theorem 2. *(Consistency) Under Assumptions 1,2,5 and either 3 or 4:*

$$\hat{\beta} \xrightarrow{p} \beta \text{ as } n \to \infty$$

[4] uniformly