cemmap
centre for microdata methods and practice

Nonseparable multinomial choice models in cross-section and panel data

Victor Chernozhukov
Iván Fernández-Val
Whitney K. Newey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP33/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Nonseparable Multinomial Choice Models in Cross-Section and Panel Data

Victor Chernozhukov* *MIT*

Iván Fernández-Val† *Boston University*

Whitney K. Newey‡ *MIT*

June 21, 2017

Abstract

Multinomial choice models are fundamental for empirical modeling of economic choices among discrete alternatives. We analyze identification of binary and multinomial choice models when the choice utilities are nonseparable in observed attributes and multidimensional unobserved heterogeneity with cross-section and panel data. We show that derivatives of choice probabilities with respect to continuous attributes are weighted averages of utility derivatives in cross-section models with exogenous heterogeneity. In the special case of random coefficient models with an independent additive effect, we further characterize that the probability derivative at zero is proportional to the population mean of the coefficients. We extend the identification results to models with endogenous heterogeneity using either a control function or panel data. In time stationary panel models with two periods, we find that differences over time of derivatives of choice probabilities identify utility derivatives "on the diagonal," i.e. when the observed attributes take the same values in the two periods. We also show that time stationarity does not identify structural derivatives "off the diagonal" both in continuous and multinomial choice panel models.

Keywords: Multinomial choice, binary choice, nonseparable model, random coefficients, panel data, control function.

*Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: vchern@mit.edu.

†Department of Economics, Boston University, Boston, MA 02215, U.S.A E-mail: ivanf@bu.edu.

‡Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: wnewey@mit.edu.

1 Introduction

Multinomial choice models are fundamental for empirical modeling of economic choices among discrete alternatives. Our starting point is the assumption that much of what determines preferences is unobserved to the econometrician. This assumption is consistent with many empirical demand and other studies where prices, income, and other observed variables explain only a small fraction of the variation in the data. From the beginning unobserved preference heterogeneity has had an important role in multinomial choice models. The classic formulation of McFadden (1974) allowed for unobserved heterogeneity through an additive term in the utility of each alternative. Hausman and Wise (1978) developed a more general specification where coefficients of regressors vary in unobserved ways among agents. Our results build on this pioneering work as well as other contributions to be discussed in what follows.

Economic theory does not generally restrict the way unobserved heterogeneity affects preferences. This observation motivates allowing for general forms of heterogeneity, as we do in this paper. We allow choice utilities to depend on observed characteristics and unobserved heterogeneity in general ways that need not be additively or multiplicatively separable. The specifications we consider allow for random coefficients but also more general specifications.

In this paper we show that derivatives of choice probabilities with respect to continuous observed attributes are weighted averages of utility derivatives. These results allow us to identify signs of utility derivatives as well as relative utility effects for different attributes. We also find that probability derivatives can be even more informative in special cases, such as random coefficients. For example, we find that for linear random coefficients with an independent additive effect the probability derivative at zero is proportional to the population mean of the coefficients.

We give choice probability derivative results for binary and multinomial choice. We do this for cross-section data where unobserved heterogeneity is independent of observed attributes. We also give derivative formulas for two cases with endogeneity. One is where the heterogeneity and utility variables are independent conditional on a control function. There we show that derivatives of choice probabilities conditional on the control function have a utility derivative interpretation. We also verify that under a common support condition, averaging over the control function gives structural function derivatives.

We also allow for endogeneity by using panel data. We give derivative formulas for discrete choice in panel data under the time stationarity condition of Manski (1987). For the constant coefficient case these give new identification results for ratios of coefficients of continuously distributed variables in panel data without requiring infinite support for any regressor or disturbance. The panel data results are partly based on Hoderlein and White (2012) as extended to the time stationary case by Chernozhukov et. al. (2015). These results use the "diagonal" where regressors in two time periods are equal to each other.

We also consider identification "off the diagonal," where regressors in different time periods

are not equal to each other. For the case of a single regressor and two time periods we construct an alternative, observationally equivalent model that is linear in the regressor. This alternative model can have a different average utility derivative off the diagonal, showing that utility average derivatives are not identified there.

The model and goal of this paper are different than that of Gautier and Kitamura (2013) and Burda, Harding, and Hausman (2008, 2010). Their goal is recover the distribution of heterogeneity in a linear random coefficients model. We consider a more general nonseparable model and a more modest goal of obtaining weighted average effects from probability derivatives. Our results provide a way of recovering certain averages of utility derivatives. Also, our results are simpler in only depending on nonparametric regressions rather than the Bayesian or deconvolution methods required to identify distributions of random coefficients.

Section 2 gives derivative formulae for binary choice. Section 3 extends these results to multinomial choice models. Section 4 obtains derivative results in the presence of a control function. Section 5 gives identification results for multinomial choice in panel data. Section 6 shows nonidentification off the diagonal. Section 7 concludes. The Appendix gives proofs.

2 Binary Choice Model

We first consider a binary choice model in cross-section data where we observe $(Y_i, X_i), (i = 1, ..., n)$ with $Y \in \{0, 1\}$ a binary choice variable and X a vector of observed characteristics (regressors). Let ε be a vector that is possibly infinite dimensional, representing unobserved aspects of agents' preferences. We will assume that the utility of choices 0 and 1 is given by $U_0(X, \varepsilon)$ and $U_1(X, \varepsilon)$ respectively. The binary choice variable Y is

$$Y = 1(U_1(X, \varepsilon) \geq U_0(X, \varepsilon)).$$

Here we impose no restrictions on the way that X and ε interact. As we will discuss, this specification includes but is not limited to random coefficient models. This specification is general enough to be like the stochastic revealed preference setting of McFadden and Richter (1991).

We begin our analysis under the assumption that X and ε are independently distributed:

Assumption 1: *(Independence) X and ε are independently distributed.*

In what follows we will relax this condition when we have a control function or when we have panel data. It is helpful to think about this model as a threshold crossing model where

$$Y = 1(\delta(X, \varepsilon) \geq 0), \quad \delta(X, \varepsilon) = U_1(X, \varepsilon) - U_0(X, \varepsilon).$$

The classic constant coefficients model is a special case where ε is a scalar and

$$\delta(X, \varepsilon) = \beta_0' X + \varepsilon.$$

This model only allows for additive unobserved heterogeneity. An important generalization is a random coefficients model where $\varepsilon = (v, \eta')'$ is a vector and

$$\delta(X, \varepsilon) = \eta' X + v.$$

This specification allows for the coefficients of the regressors to vary with the individual. Hausman and Wise (1978) proposed such a specification for Gaussian ε. Berry, Levinsohn, and Pakes (1995, BLP henceforth) proposed a mixed logit/Gaussian specification where v is the difference of Type I extreme value variables and η is Gaussian. Gautier and Kitamura (2013) gave results on identification and estimation of the distribution of η when that distribution is unknown. The nonseparable specification we consider is more general in allowing for $\delta(X, \varepsilon)$ to be nonlinear in X and/or ε.

In this binary choice setting the choice probability is given by

$$P(X) := \Pr(Y = 1 \mid X) = \Pr(\delta(X, \varepsilon) \geq 0 \mid X) = \int 1(\delta(X, \varepsilon) \geq 0) F_\varepsilon(d\varepsilon),$$

where F_ε denotes the CDF of ε. Here we derive a formula that relates the derivatives of the choice probability with respect to X to the derivatives of $\delta(X, \varepsilon)$. Let ∂_x denote the vector of partial derivatives with respect to all the continuously distributed components of X and ∂_v the partial derivative with respect to a scalar v.

Assumption 2: *(Monotonicity) For some η and v, $\varepsilon = (\eta', v)'$ where v is a scalar, $\delta(x, \varepsilon) = \delta(x, \eta, v)$ is continuously differentiable in x and v, and there is $C > 0$ such that $\partial_v \delta(x, \eta, v) \geq 1/C$ and $\|\partial_x \delta(x, \eta, v)\| \leq C$ everywhere. The variable v is continuously distributed conditional on η with a conditional density $f_v(v \mid \eta)$ that is bounded and continuous in v.*

As discussed below, for binary choice this condition will be equivalent to $\delta(x, \varepsilon)$ being additive in v that is continuously distributed with a density satisfying the above condition. Let $f_{\delta(x,\varepsilon)}$ denote the density of $\delta(x, \varepsilon)$.

Theorem 1: *If Assumptions 1 and 2 are satisfied then,*

$$\partial_x P(x) = \mathrm{E}[\partial_x \delta(x, \varepsilon) \mid \delta(x, \varepsilon) = 0] \cdot f_{\delta(x,\varepsilon)}(0).$$

Theorem 1 shows that derivatives of the choice probability are scalar multiples of averages of the derivative $\partial_x \delta(x, \varepsilon)$ conditional on being at the zero threshold, i.e. conditional on being

indifferent between the two choices. Here the choice probability is one minus the CDF of $\delta(x, \varepsilon)$ at zero, so that the choice probability derivative is the negative of the CDF derivative at zero, i.e.

$$\partial_x P(x) = -\partial_x \Pr(\delta(x, \varepsilon) \leq y)|_{y=0}.$$

The formula in Theorem 1 corresponds to the derivative of the CDF of a nonseparable model derived in Blomquist et. al. (2014), which builds on the quantile derivative result of Hoderlein and Mammen (2007). The conclusion of Theorem 1 is an important application of this formula to the choice probability derivative in a nonseparable model.

Assumption 2 restricts our model somewhat relative to the stochastic revealed preference model of McFadden and Richter (1991). It is possible to obtain another informative derivative formula under regularity conditions like those of Sasaki (2015) and Chernozhukov, Fernandez-Val, and Luo (2015), that are different than Assumption 2. Those conditions lead to a more general formula for $\partial_x P(x)$. That formula allows for multiple crossings of the threshold 0 while the monotonicity condition in Assumption 2 implies that there is only one threshold crossing conditional on η. It is not clear how restrictive Assumption 2 or the alternative conditions are relative to the stochastic revealed preference setting of McFadden and Richter (1991). For brevity we omit further discussion of this issue.

Another special case of a nonseparable model is an index model where $\delta(x, \varepsilon) = h(\beta_0' x, \varepsilon)$ for some constant coefficients β_0. Here $P(X) = \tau(\beta_0' X)$ for $\tau(u) = \Pr(h(u, \varepsilon) \geq 0) = \int 1(h(u, \varepsilon) \geq 0) F_\varepsilon(d\varepsilon)$. This model results in a choice probability that depends on X only through the index $\beta_0' X$, similarly to Stoker (1986), Ichimura (1993), and Ai (1997). By Theorem 1 it follows that

$$\tau_u(u) = \partial_u \tau(u) = \mathrm{E}[\partial_u h(u, \varepsilon) \mid h(u, \varepsilon) = 0] \cdot f_{h(u,\varepsilon)}(0).$$

Differentiating with respect to the continuous components of X gives the well known index derivative formula,

$$\partial_x P(x) = \beta_0 \cdot \tau_u(\beta_0' x).$$

Here the derivatives of the choice probability are scalar multiples of the components of β_0.

There is an alternative version of Theorem 1 that provides further insight and motivates our multinomial choice results that follow in Section 3. By the monotonicity condition of Assumption 2

$$\delta(x, \eta, v) \geq 0 \iff h(x, \eta) + v \geq 0,$$

where $h(x, \eta) := -\delta^{-1}(x, \eta, r)|_{r=0}$ and the function inverse is with respect to the v argument in $\delta(x, \eta, v)$. Then the choice probability is

$$P(x) = \mathrm{E}[\Pr(v \geq -h(x, \eta) \mid \eta)] = \mathrm{E}[1 - F_v(-h(x, \eta) \mid \eta)] = \int [1 - F_v(-h(x, \eta) \mid \eta)] F_\eta(d\eta), \tag{2.1}$$

where $F_v(v \mid \eta)$ is the conditional CDF of v given η, and $F_\eta(\eta)$ is the CDF of η.

This model is also observationally equivalent to a threshold crossing model

$$Y = 1(\tilde{\delta}(x, \eta, \tilde{v}) \geq 0), \quad \tilde{\delta}(x, \eta, \tilde{v}) := \tilde{h}(x, \eta) + \tilde{v},$$

with a disturbance $\tilde{v}$ that is independent of X and η. Let $\tilde{h}(X, \eta) := -F_v(-h(X, \eta) \mid \eta)$ and $\tilde{v} \sim U(0, 1)$ independently of (X, η). Then since $0 \leq -\tilde{h}(x, \eta) \leq 1$,

$$\Pr(\tilde{v} \geq -\tilde{h}(x, \eta)) = \mathrm{E}[1 - \Pr(\tilde{v} \leq -\tilde{h}(x, \eta) \mid \eta)] = \mathrm{E}[1 + \tilde{h}(x, \eta)] = P(x).$$

Moreover, by Theorem 1,

$$\mathrm{E}[\partial_x \tilde{\delta}(x, \eta, \tilde{v}) \mid \tilde{\delta}(x, \eta, \tilde{v}) = 0] \cdot f_{\tilde{\delta}(x,\eta,\tilde{v})}(0) = \mathrm{E}[\partial_x \delta(x, \eta, v) \mid \delta(x, \eta, v) = 0] \cdot f_{\delta(x,\eta,v)}(0).$$

In this sense there is no loss of generality in assuming that v is independent of (X, η).

Differentiating the expression of $P(x)$ in (2.1) with respect to x gives the following result:

COROLLARY 2: *If Assumptions 1 and 2 are satisfied then*

$$\partial_x P(x) = \mathrm{E}[\{\partial_x h(x, \eta)\} f_v(-h(x, \eta) \mid \eta)] = \int [\partial_x h(x, \eta)] f_v(-h(x, \eta) \mid \eta) dF_\eta(\eta).$$

The formula given here is relatively easy to interpret. We can clearly see that the derivative of the choice probability is a weighted average of the derivative $\partial_x h(x, \eta)$ where the weight is the conditional pdf of v given η evaluated at $-h(x, \eta)$. For instance, consider a binary mixed logit specification like BLP where v is the difference of two Type I extreme value disturbances that are independent of η and each other, and where η are random coefficients with $h(x, \eta) = \eta' x$, so that we can take $\delta(X, \varepsilon) = \eta' x + v$. Let $f_v(v) = e^v/[1 + e^v]^2$ be the logit pdf. Then from the previous formula,

$$\partial_x P(x) = \mathrm{E}[f_v(-\eta' x)\eta].$$

Here the probability derivative is a weighted average of the random coefficients, with the weight being the logit pdf values evaluated at the regression $-\eta' x$.

It is interesting to note that at $x = 0$ the conclusion of Corollary 2 implies

$$\partial_x P(x)|_{x=0} = \mathrm{E}[f_v(0)\eta] = f_v(0) \cdot \mathrm{E}[\eta]. \tag{2.2}$$

Thus, when X has positive density around zero and v and η are independent, the derivative of the choice probability at zero estimates the expected value of the random coefficients up to scale. Consequently

$$\left[\partial_{x_j} P(x)/\partial_{x_k} P(x)\right]\big|_{x=0} = \mathrm{E}[\eta_j]/\mathrm{E}[\eta_k].$$

This equation is a binary choice analog of the result that in a linear random coefficients model the regression of y on X estimates the expectation of the coefficients. With binary choice only ratios of coefficients are identified, so here only ratios of expected values are identified.

COROLLARY 3: *If Assumptions 1 and 2 are satisfied, $\delta(x,\varepsilon) = \eta' x + v$, and v is independent of η, then equation (2.2) is satisfied.*

Weighted average derivatives of the choice probability can be used to summarize the effect of x on $h(x,\eta)$. From Theorem 1 we can see that weighted average derivatives will be weighted averages of $\partial_x h(x,\eta)$ conditional on $\delta(x,\varepsilon) = 0$. In particular, for any bounded nonnegative function $w(x)$ it follows from Corollary 2 that

$$\mathrm{E}[w(X)\partial_x P(X)] = \mathrm{E}[w(X) f_v(-h(X,\eta) \mid \eta) \{\partial_x h(X,\eta)\}].$$

Here the derivative is weighted by both $w(X)$ and the density $f_v(-h(X,\eta) \mid \eta)$. The density weight is present because the derivatives of $h(x,\eta)$ have been "filtered" through the discrete choice and so the probability derivative only recovers effects where $h(x,\eta) + v = 0$.

3 Multinomial Choice Models

In this Section we extend the analysis to the nonseparable multinomial choice model. Here there are J choices $j = 1, ..., J$. Each choice has a utility $U_j(X,\varepsilon)$ associated with it, depending on observed characteristics X and unobserved characteristics ε. Let Y_j denote the choice indicator that is equal to one if the j^{th} alternative is chosen and zero otherwise. Then

$$Y_j = 1(\{U_j(X,\varepsilon) \geq U_k(X,\varepsilon); k = 1, ..., J\}).$$

The probability $P_j(x) := \Pr(Y_j = 1 \mid X = x)$ that j is chosen conditional on $X = x$ is the probability that $U_j(x,\varepsilon)$ is the maximum utility among the J choices, i.e.

$$P_j(x) = \Pr\{U_j(x,\varepsilon) \geq U_k(x,\varepsilon); k = 1, ..., J\} = \int 1(\{U_j(x,\varepsilon) \geq U_k(x,\varepsilon); k = 1, ..., J\}) F_\varepsilon(d\varepsilon),$$

where we maintain Assumption 1 and assume that the probability of ties is zero.

We can obtain a useful formula for the derivative of this probability under a condition analogous to Assumption 2. Recall that the monotonicity condition of Assumption 2 is equivalent to the existence of a scalar additive disturbance. Here we will impose scalar additive disturbances from the outset.

ASSUMPTION 3: *(Multinomial Choice) There are $\eta, v_j, u_j(x,\eta), (j = 1, ..., J)$ such that $\varepsilon = (\eta', v')'$ for $v := (v_1, ..., v_J)$,*

$$U_j(x,\varepsilon) = u_j(x,\eta) + v_j,$$

and $u_j(x,\eta)$ is continuously differentiable in x with bounded derivative.

In this condition we assume directly an additive disturbance condition that we showed is equivalent to Assumption 2 in the binary case. Assumption 3 generalizes that additive disturbance condition to multinomial choice. Similarly to binary choice, we are not sure what restrictions this additive specification would impose in the stochastic revealed preference setting of McFadden and Richter (1991).

As for binomial choice we could formulate the results in terms of differences of utilities. However, we find it convenient to work directly with choice specific utilities $U_j(x,\varepsilon) = u_j(x,\eta) + v_j$ rather than differences. Let $u := (u_1, ..., u_J)$ denote a $J \times 1$ vector of constants and

$$p_j(u \mid \eta) := \Pr(u_j + v_j \geq u_k + v_k; k = 1, ..., J \mid \eta).$$

This $p_j(u \mid \eta)$ is just the usual multinomial choice probability, conditioned on η. For example, if v is a vector of i.i.d. Type I extreme value random variables independent of η, then p_j has the multinomial logit form

$$p_j(u \mid \eta) = \frac{e^{u_j}}{\sum_{k=1}^{J} e^{u_k}}.$$

In general when $f_v(v \mid \eta)$ is continuous, $p_j(u \mid \eta)$ will be continuously differentiable in each u_k. Let

$$p_{jk}(u \mid \eta) := \partial p_j(u \mid \eta)/\partial u_k, \quad u(x,\eta) := (u_1(x,\eta), ..., u_J(x,\eta))'.$$

THEOREM 4: *If Assumptions 1 and 3 are satisfied, the conditional density $f_v(v \mid \eta)$ of v given η is continuous in v,* and *$p_{jk}(u \mid \eta), (j,k = 1, ..., J)$ are bounded, then $P_j(x)$ is differentiable in x and*

$$\partial_x P_j(x) = \mathrm{E}\left[\sum_{k=1}^{J} p_{jk}(u(x,\eta) \mid \eta)\partial_x u_k(x,\eta)\right] = \int \left[\sum_{k=1}^{J} p_{jk}(u(x,\eta) \mid \eta)\partial_x u_k(x,\eta)\right] F_\eta(d\eta).$$

As an example consider again the multinomial logit where v consists of i.i.d Type I extreme value random variables that are independent of η. Define $\tilde{p}_j(x,\eta) := e^{u_j(x,\eta)} / \sum_{k=1}^{J} e^{u_k(x,\eta)}$. Then,

$$\partial_x P_j(x) = \mathrm{E}\left[\tilde{p}_j(x,\eta)\left\{\partial_x u_j(x,\eta) - \sum_{k=1}^{J} \tilde{p}_k(x,\eta)\partial_x u_k(x,\eta)\right\}\right].$$

For example, if some x^ℓ affects only $u_{j_\ell}(x,\eta)$ for some j_ℓ, then

$$\partial_{x^\ell} P_{j_\ell}(x) = \int \tilde{p}_{j_\ell}(x,\eta)\left\{1 - \tilde{p}_{j_\ell}(x,\eta)\right\}\partial_{x^\ell} u_{j_\ell}(x,\eta) F_\eta(d\eta).$$

An important class of examples are those where $u_j(x,\eta) = \eta' x^j$ for choice specific observable characteristics x^j. This example is similar to BLP where x^j could be thought of as the characteristics of an object for choice j, such as characteristics of the j^{th} car type. Here an additional unit of some component of x^j affects the utility the same for each alternative j, i.e. there is no utility interaction between the choice j and the characteristics x^j. In this class of examples,

$$\partial_{x^k} P_j(x) = \mathrm{E}[p_{jk}(u(x,\eta) \mid \eta)\eta].$$

Here we see that the derivative of the j^{th} choice probability with respect to the regressor vector x^k for the k^{th} alternative is an expectation of the random coefficients multiplied by a scalar $\partial p_j(u(x,\eta) \mid \eta)/\partial u_k$. As in the binary case if η is a constant vector β_0 then

$$\partial_{x^k} P_j(x) = \mathrm{E}[p_{jk}(u(x,\eta) \mid \eta)] \cdot \beta_0,$$

so that the derivative of the choice probability is proportional to β_0 for all x^k. Also if v is independent of η so that $p_j(u \mid \eta) = p_j(u)$, and each of the characteristic vectors is zero, then the scalar is a constant and

$$\partial_{x^k} P_j(x)|_{x^1=\cdots=x^J=0} = p_{jk}(u)|_{u=0} \cdot \mathrm{E}[\eta].$$

Similarly to the binary case the derivative of the probability at the origin is a scalar multiple of the expectation of the random coefficients.

4 Control Functions

A model where it is possible to allow for nonindependence between ε and X is one where there is an observable or estimable control function w satisfying

Assumption 4: *(Control Function) X and ε are independently distributed conditional on w.*

As shown in Blundell and Powell (2004) and Imbens and Newey (2009), conditioning on a control function helps to identify objects of interest.[1] Here we show how a control function can be used to estimate averages of utility derivatives. These derivatives will be exactly analogous to those considered previously, except that we also condition on the control function.

[1] Berry and Haile (2010) considered an alternative approach based on the availability of "special regressors" and instrumental variables satisfying completeness conditions in multinomial choice demand models where the endogenous part of the unobserved heterogeneity is scalar. This approach identifies the entire distribution of random utilities.

The choices Y_j are determined as before but now we consider choice probabilities that condition on w as well as X. These probabilities are given by

$$P_j(X, w) := \Pr(Y_j = 1 \mid X, w).$$

Let $u := (u_1, ..., u_J)$ denote a $J \times 1$ vector of constants and

$$p_j(u \mid \eta, w) := \Pr(u_j + v_j \geq u_k + v_k; k = 1, ..., J \mid \eta, w).$$

This $p_j(u \mid \eta, w)$ is just the usual multinomial choice probability, conditioned on η and w. For example, if v is a vector of i.i.d. Type I extreme value random variables independent of η and w, then p_j has the multinomial logit form

$$p_j(u \mid \eta, w) = \frac{e^{u_j}}{\sum_{k=1}^{J} e^{u_k}}.$$

In general, when the conditional density of v given η and w is continuous, $p_j(u \mid \eta, w)$ will be continuously differentiable in each u_k. Let $p_{jk}(u \mid \eta, w) := \partial p_j(u \mid \eta, w)/\partial u_k$ and $u(x, \eta) := (u_1(x, \eta), ..., u_J(x, \eta))'$ as before.

THEOREM 5: *If Assumptions 3 and 4 are satisfied, the conditional density $f_v(v \mid \eta, w)$ of v given η and w is continuous in v, and $p_{jk}(u \mid \eta, w), (j, k = 1, ..., J)$ are bounded, then $P_j(x, w)$ is differentiable in x and*

$$\begin{aligned}\partial_x P_j(x, w) &= \mathrm{E}\left[\sum_{k=1}^{J} p_{jk}(u(x, \eta) \mid \eta, w)\partial_x u_k(x, \eta) \mid w\right] \\ &= \int \left[\sum_{k=1}^{J} p_{jk}(u(x, \eta) \mid \eta, w)\partial_x u_k(x, \eta)\right] F_\eta(d\eta \mid w).\end{aligned}$$

As an example consider again multinomial logit where v consists of i.i.d Type I extreme value random variables that are independent of η and w. Define $\tilde{p}_j(x, \eta) := e^{u_j(x,\eta)}/\sum_{k=1}^{J} e^{u_k(x,\eta)}$. Then,

$$\begin{aligned}\partial_x P_j(x, w) &= \mathrm{E}\left[\tilde{p}_j(x, \eta)\left\{\partial_x u_j(x, \eta) - \sum_{k=1}^{J} \tilde{p}_k(x, \eta)\partial_x u_k(x, \eta)\right\} \mid w\right] \\ &= \int \tilde{p}_j(x, \eta)\left\{\partial_x u_j(x, \eta) - \sum_{k=1}^{J} \tilde{p}_k(x, \eta)\partial_x u_k(x, \eta)\right\} F(d\eta \mid w).\end{aligned}$$

For example, if some x^ℓ affects only $u_{j_\ell}(x, \eta)$ for some j_ℓ, then

$$\partial_{x^\ell} P_{j_\ell}(x, w) = \int \tilde{p}_{j_\ell}(x, \eta)\{1 - \tilde{p}_{j_\ell}(x, \eta)\}\partial_{x^\ell} u_{j_\ell}(x, \eta) F(d\eta \mid w).$$

An important class of examples are those where $u_j(x,\eta) = \eta' x^j$ for choice specific observable characteristics x^j. This example is similar to BLP where x^j could be thought of as the characteristics of an object for choice j, such as characteristics of the j^{th} car type. Here an additional unit of some component of x^j affects the utility the same for each alternative j, i.e. there is no utility interaction between the choice j and the characteristics x^j. In this class of examples,

$$\partial_{x^k} P_j(x,w) = \int [p_{jk}(u(x,\eta) \mid \eta, w) \cdot \eta] F(d\eta \mid w).$$

Here we see that the derivative of the j^{th} choice probability with respect to the vector x^k for the k^{th} alternative is an expectation of the random coefficients multiplied by a scalar $\partial p_j(u(x,\eta) \mid \eta, w)/\partial u_k$. As above, if η is a constant vector β_0, then

$$\partial_{x^k} P_j(x,w) = \beta_0 \cdot \int p_{jk}(u(x,\eta) \mid \eta, w) F(d\eta \mid w),$$

so that the derivative of the choice probability is proportional to β_0 for all x and w. Also, if v is independent of η and w so that $p_j(u \mid \eta, w) = p_j(u)$, then at $x = 0$ the scalar is a constant and

$$\partial_{x^k} P_j(x,w)|_{x^1=\cdots=x^J=0} = p_{jk}(u)|_{u=0} \cdot \mathrm{E}[\eta \mid w].$$

Similarly to above, the derivative of the probability at the origin is a scalar multiple of the expectation of the random coefficients, here conditional on the control function.

As is known from the previous literature, integrating over the marginal distribution of the control function gives probability derivatives identical to those for X and ε independent, when a common support condition is satisfied:

COROLLARY 6: *If Assumptions 3 and 4 are satisfied, the conditional density $f_v(v \mid \eta, w)$ of v given η and w is continuous in v and bounded, and the conditional support for w* given $X = x$ *equals the marginal support for w, then $P_j(x,w)$ is differentiable in x and*

$$\int \partial_x P_j(x,w) F_w(dw) = \int \left[\sum_{k=1}^{J} p_{jk}(u(x,\eta) \mid \eta) \partial_x u_k(x,\eta) \right] F_\eta(d\eta),$$

where $F_w(w)$ is the CDF of w.

It is interesting to note that the common support condition is not needed for identification of interesting effects. Averages of utility derivatives are identified from probability derivatives, conditional on the control function, as in Theorem 5. Also, because η is independent of X conditional on w, averages over η conditional on X can be identified by integrating the objects in Theorem 5 over the conditional distribution of w given X. This integration gives local average probability responses analogous to the local average response given in Altonji and Matzkin (2005). In addition, averaging over the joint distribution of X and w gives average derivatives analogous to those considered by Imbens and Newey (2009). None of these effects rely on the common support condition.

5 Panel Data

Panel data can also help us identify averages of utility derivatives when X and ε are not independent. Invariance over time of the distribution of ε conditional on the observed X for all time periods can allow us to identify utility derivative averages analogous to those we have considered. This invariance over time of the distribution of ε conditional on regressors is the basis of previous panel identification results by Manski (1987), Honore (1992), Abrevaya (2000), Chernozhukov et. al. (2013), Graham and Powell (2012), Chernozhukov et. al. (2015), and is an important hypothesis in Hoderlein and White (2012). Pakes and Porter (2014) and Shi et. al. (2017) have given identification results for multinomial choice models under this condition. These papers allow for some time effects while Evdokimov (2010) allowed for general time effects while imposing independence and additivity among disturbances.

In this Section we consider a panel multinomial choice model with panel binary choice as a special case. For simplicity we just consider two time periods. It is straightforward to extend the results to more than two time periods. Let Y_{jt} denote the choice indicator, equal to 1 if alternative j is chosen in time period t. Let X_t denote the observable characteristics and ε_t the unobservable ones in period t. We assume that choice is based on a time stationary utility function $U_j(x, \varepsilon) = u_j(x, \eta) + v_j$ having the additive form considered in the previous Section. The time stationarity of the utility is important to our results though it may be possible to relax that condition similarly to Chernozhukov et. al. (2015).

It is assumed that the individual makes the choice that maximizes utility in each time period, so that

$$Y_{jt} = 1(\{U_j(X_t, \varepsilon_t) \geq U_k(X_t, \varepsilon_t), k = 1, ..., J\}), (j = 1, ..., J, t = 1, 2).$$

To identify averages of utility derivatives we use choice probabilities conditional on the regressors $\mathbf{X} := (X_1, X_2)$ for both time periods, given by

$$P_{jt}(\mathbf{X}) := \Pr(Y_{jt} = 1 \mid X_1, X_2).$$

With panel data we can replace the assumption of independence of ε and $\mathbf{X}$ with the following time stationarity condition:

ASSUMPTION 5: *(Time Stationarity) The distribution of ε_t given $\mathbf{X} = (X_1, X_2)$ does not depend on t.*

For a constant vector $u := (u_1, ..., u_J)$ let

$$p_j(u \mid \eta_t, \mathbf{X}) := \Pr(u_j + v_{tj} \geq u_k + v_{tk}; k = 1, ..., J \mid \eta_t, \mathbf{X}).$$

This is like the usual choice probability, as discussed earlier, only now it depends on $\mathbf{X}$ as well as η_t. What allows us to identify derivative effects despite the dependence of p_j on $\mathbf{X}$ is that p_j does

not depend on t because of the time stationarity condition of Assumption 5. Time stationarity allows us to difference out the confounding effect of $\mathbf{X}$ that acts through the correlation of $\mathbf{X}$ with ε. By iterated expectations the choice probability is

$$P_{jt}(\mathbf{X}) = \Pr(Y_{jt} = 1 \mid \mathbf{X}) = \mathrm{E}[p_j(u(X_t, \eta_t) \mid \eta_t, \mathbf{X}) \mid \mathbf{X}].$$

The difference over two time periods is

$$P_{j2}(\mathbf{X}) - P_{j1}(\mathbf{X}) = \mathrm{E}[p_j(u(X_2, \eta_2) \mid \eta_2, \mathbf{X}) \mid \mathbf{X}] - \mathrm{E}[p_j(u(X_1, \eta_2) \mid \eta_2, \mathbf{X}) \mid \mathbf{X}],$$

where we have used the time stationarity in replacing η_1 by η_2 in $P_{j1}(\mathbf{X})$. When we differentiate this with respect to X_2 the presence of $P_{j1}(\mathbf{X})$ removes all the derivatives with respect to X_2 except the utility derivatives, where $X_1 = X_2$.

If the conditional density $f_v(v \mid \eta_t, \mathbf{X})$ is continuous in v then $p_j(u \mid \eta_t, \mathbf{X})$ will be continuously differentiable in u. Let $p_{jk}(u \mid \eta_t, \mathbf{X}) := \partial p_j(u \mid \eta_t, \mathbf{X})/\partial u_k$ and $\mu_j(x \mid \mathbf{X}) := \mathrm{E}[p_j(u(x, \eta_t) \mid \eta_t, \mathbf{X}) \mid \mathbf{X}]$.

THEOREM 7: *If Assumptions 3 and 5 are satisfied, the conditional density $f_v(v \mid \eta_t, \mathbf{X})$ of v_t given η_t and $\mathbf{X}$ is continuous in v, $p_{jk}(u \mid \eta_t, \mathbf{X}), (j, k = 1, ..., J)$ are bounded, and $\mu_j(x \mid X_1, X_2)$ is differentiable in x and X_2 then,*

$$\begin{aligned}\partial_{X_2} \mathrm{E}[Y_{j2} - Y_{j1} \mid \mathbf{X}]|_{X_1 = X_2} &= \partial_{X_2}\{P_{j2}(\mathbf{X}) - P_{j1}(\mathbf{X})\}|_{X_1 = X_2} \\ &= \mathrm{E}\left[\sum_{k=1}^{J} p_{jk}(u(X_2, \eta_2) \mid \eta_2, \mathbf{X}) \partial_x u_k(X_2, \eta_2) \mid \mathbf{X}\right]\Bigg|_{X_1 = X_2}.\end{aligned}$$

As an example consider again multinomial logit where v_t consists of i.i.d Type I extreme value random variables that are independent of η_t and of $\mathbf{X}$. Define $\tilde{p}_j(x, \eta) := e^{u_j(x,\eta)} / \sum_{k=1}^{J} e^{u_k(x,\eta)}$. Then,

$$\begin{aligned}&\partial_{X_2} \mathrm{E}[Y_{j2} - Y_{j1} \mid \mathbf{X}]|_{X_1 = X_2} \\ &\qquad = \mathrm{E}\left[\tilde{p}_j(X_2, \eta_2)\left\{\partial_x u_j(X_2, \eta_2) - \sum_{k=1}^{J} \tilde{p}_k(X_2, \eta_2) \partial_x u_k(X_2, \eta_2)\right\} \mid \mathbf{X}\right]\Bigg|_{X_1 = X_2}.\end{aligned}$$

For example, if some X_2^ℓ affects only $u_{j_\ell}(X_2, \eta)$, then for logit

$$\partial_{X_2^\ell} \mathrm{E}[Y_{j_\ell 2} - Y_{j_\ell 1} \mid \mathbf{X}]\Big|_{X_1 = X_2} = \mathrm{E}[\tilde{p}_{j_\ell}(X_2, \eta_2)\{1 - \tilde{p}_{j_\ell}(X_2, \eta_2)\} \partial_{x^\ell} u_{j_\ell}(X_2, \eta_2) \mid \mathbf{X}]|_{X_1 = X_2}.$$

An important class of examples are those where $u_j(x, \eta) = \eta' x^j$ for choice specific observable characteristics x^j. This example is like panel BLP where x^j could be thought of as the characteristics of an object for choice j, such as characteristics of the j^{th} car type. Here an additional

unit of some component of x^j affects the utility the same for each alternative j, i.e. there is no utility interaction between the choice j and the characteristics x^j. In this class of examples,

$$\partial_{X_2^k} \mathrm{E}[Y_{j2} - Y_{j1} \mid \mathbf{X}]\Big|_{X_1=X_2} = \mathrm{E}[p_{jk}(u(X_2, \eta_2) \mid \eta_2, \mathbf{X}) \cdot \eta_2 \mid \mathbf{X}]|_{X_1=X_2}.$$

Here we see that the derivative of the j^{th} choice probability difference with respect to X_2^k for the k^{th} alternative is an expectation of the random coefficients multiplied by a scalar $p_{jk}(u(X_2, \eta) \mid \eta)$.

Time stationary panel data provides a way of controlling for endogeneity of prices in imperfectly competitive markets where the price is one element of X_t. The time stationarity condition of Assumption 5 allows for unobserved features of preferences corresponding to ε_t to be correlated with $\mathbf{X}$ in unspecified ways, as long as that relationship is the same for each time period. In particular, as mentioned earlier, components of ε that do not vary over time automatically satisfy this condition. In this sense Assumption 5 is a very general condition for preferences that do not vary over time. It can also be extended to settings where the dimension t corresponds to different markets or locations instead of time periods.

Similar to the cross section case, if η is a constant vector β_0 then

$$\partial_{X_2^k} \mathrm{E}[Y_{j2} - Y_{j1} \mid \mathbf{X}]\Big|_{X_1=X_2} = \mathrm{E}[p_{jk}(u(X_2, \eta_2) \mid \eta_2, \mathbf{X}) \mid \mathbf{X}]|_{X_1=X_2} \cdot \beta_0. \tag{5.1}$$

Thus we find that that the derivative of the choice probability is proportional to β_0 for all $X_1 = X_2$ in a panel data multinomial choice model where $u_j(x, \eta) = \beta_0' x^j$.

THEOREM 8: *If Assumption 4 is satisfied, $U_j(x, \varepsilon) = \beta_0' x^j + v_j$, and $\mu_j(x \mid X_1, X_2)$ is differentiable in x and X_2, then for each j and k,* equation (5.1) is satisfied. *Also, if $\mathrm{E}[p_{jk}(u(X_1, \eta_t) \mid \eta_t, \mathbf{X}) \mid \mathbf{X}]|_{X_1=X_2} \neq 0$ for some j, k, and X_1, then β_0 is identified up to scale.*

This gives an identification result for multinomial choice models in panel data. It shows that the vector of coefficients of continuous regressors in a multinomial choice model with additive fixed effect is identified up to scale from the diagonal where $X_1 = X_2$. This identification result holds even if X_t is bounded, unlike that of Manski (1987). It can also hold even with v_t having bounded support, unlike that of Shi et. al. (2017). Note though that we only identify the coefficients β_0 up to scale, whereas the identification results of Manski (1987) and Shi et. al. (2017) are for β_0 itself. In independent work Chen and Wang (2017) has recently shown that in panel binary choice the entire vector β_0 can be identified up to scale if just one component of X_t is continuously distributed.

6 Nonidentification Off the Diagonal

The panel data results show identification of utility derivatives on the diagonal where $X_1 = X_2$. We can also show that off the diagonal, where $X_1 \neq X_2$, utility derivatives are not identified. Specifically, off the diagonal one can obtain multiple values of conditional expectations of utility derivatives from the same distribution of the data.

To provide intuition we first show nonidentification for the smooth case where

$$Y_t = \phi(X_t, \varepsilon_t), \tag{6.1}$$

X_t is a scalar, $\phi(x, \varepsilon)$ is continuously differentiable in x, and the distribution of ε_t given $\mathbf{X} = (X_1, X_2)'$ is time stationary. Suppose that equation (6.1) is true. We can construct an alternative, observationally equivalent nonseparable model with time stationary disturbances as

$$\begin{aligned} &Y_t = \tilde{\varepsilon}_a + \tilde{\varepsilon}_b X_t = \tilde{\phi}(X_t, \tilde{\varepsilon}), \quad \tilde{\phi}\left(x, \tilde{\varepsilon}\right) := \tilde{\varepsilon}_a + \tilde{\varepsilon}_b x, \quad \tilde{\varepsilon} := (\tilde{\varepsilon}_a, \tilde{\varepsilon}_b)', \\ &\tilde{\varepsilon}_a := Y_1 - \tilde{\varepsilon}_b X_1, \quad \tilde{\varepsilon}_b := \left(Y_2 - Y_1\right) / (X_2 - X_1). \end{aligned}$$

By construction $\tilde{\varepsilon}$ does not vary with t, so that it is time stationary. Also, $Y_t = \tilde{\phi}(X_t, \tilde{\varepsilon})$ so that the alternative model is observationally equivalent to the original one. Furthermore, the expected value of $\tilde{\phi}_x\left(X_2, \tilde{\varepsilon}\right) := \partial \tilde{\phi}(x, \tilde{\varepsilon}) / \partial x|_{x=X_2}$ conditional on $\mathbf{X}$ is

$$\mathrm{E}[\tilde{\phi}_x\left(X_2, \tilde{\varepsilon}\right) \mid \mathbf{X}] = \mathrm{E}[\tilde{\varepsilon}_b \mid \mathbf{X}] = \frac{\mathrm{E}[\phi(X_2, \varepsilon_2) - \phi(X_1, \varepsilon_2) \mid \mathbf{X}]}{X_2 - X_1.}. \tag{6.2}$$

In contrast

$$\mathrm{E}[\phi_x\left(X_2, \varepsilon_t\right) \mid \mathbf{X}] = \mathrm{E}\left[\frac{\partial \phi(X_2, \varepsilon_2)}{\partial x} \mid \mathbf{X}\right]. \tag{6.3}$$

In general the expected derivative in equation (6.2) will not equal the expected derivative in equation (6.3) when $\mathrm{E}[\phi(x, \varepsilon_2) \mid \mathbf{X}]$ is nonlinear in x over the set where $X_1 \neq X_2$. Thus we have constructed an observationally equivalent nonseparable model with $\mathrm{E}[\tilde{\phi}_x\left(X_2, \tilde{\varepsilon}\right) \mid \mathbf{X}] \neq \mathrm{E}[\phi_x\left(X_2, \varepsilon_2\right) \mid \mathbf{X}]$, implying that $\mathrm{E}[\phi_x\left(X_2, \varepsilon_2\right) \mid \mathbf{X}]$ is not identified, on the set where $X_1 \neq X_2$. The following is a precise statement of this nonidentification result.

THEOREM 9: *If i)* $Y_t = \phi(X_t, \varepsilon_t)$, ε_t *is time stationary conditional on* $\mathbf{X}$; *ii)* $\phi(x, \varepsilon)$ *is continuously differentiable in* x *with bounded derivative; and iii) there does not exist any* $\Delta(\mathbf{X})$ *such that*

$$\mathrm{E}[\phi(X_2, \varepsilon_2) - \phi(X_1, \varepsilon_2) \mid \mathbf{X}] = \Delta(\mathbf{X})(X_2 - X_1)$$

for $X_1 \neq X_2$, *then* $\mathrm{E}[\phi_x\left(X_2, \varepsilon_2\right) \mid \mathbf{X}]$ *is not identified on the set* $X_1 \neq X_2$.

It is interesting that the form of the alternative, observationally equivalent model $Y_t = \tilde{\varepsilon}_a + \tilde{\varepsilon}_b X_t$ is linear in X_t. This is the model considered by Graham and Powell (2012). Observational

equivalence of this model to the true model means that it is impossible to distinguish from the data a linear in x model from a nonlinear one, when there is one regressor and two time periods. Furthermore, this also shows that the object estimated by the Graham and Powell (2012) estimator will be the expected difference quotient

$$\mathrm{E}[\tilde{\varepsilon}_b] = \mathrm{E}\left[\frac{\phi(X_2, \varepsilon_2) - \phi(X_1, \varepsilon_2)}{X_2 - X_1}\right].$$

This could be an interesting object. Of course one might also be interested in the expected derivative on the diagonal given by $\partial_{X_2}\mathrm{E}[Y_2 - Y_1 \mid \mathbf{X}]|_{X_1=X_2}$; see Hoderlein and White (2012) and Chernozhukov et. al. (2015). It might be best to report both kinds of effects in practice, given the impossibility of distinguishing a linear from a nonlinear model when there is a scalar X_t and two time periods.

We can give an analogous result for binary choice. Consider a binary choice panel model where $P_t(\mathbf{X}) = \mathrm{E}[1 - F_v(-h(X_t, \eta_t) \mid \eta_t, \mathbf{X}) \mid \mathbf{X}]$, similar to the cross-section model in (2.1). Consider

$$\begin{aligned}
&\tilde{Y}_t = 1(\tilde{\eta}_a + \tilde{\eta}_b X_t - 1 + \tilde{v}_t \geq 0), \tilde{v}_t \sim U(0,1), \\
&\tilde{\eta}_a := P_1(\mathbf{X}) - \tilde{\eta}_b X_1, \tilde{\eta}_b := [P_2(\mathbf{X}) - P_1(\mathbf{X})]/(X_2 - X_1),
\end{aligned}$$

where $\tilde{v}_t$ is independent of everything else. Note that $\Pr(\tilde{Y}_t = 1 \mid \mathbf{X}) = P_t(\mathbf{X}), (t = 1, 2)$, so that this alternative model is equivalent to the first one. In this alternative model $F_{\tilde{v}}(h \mid \tilde{\eta}, \mathbf{X}) = h$, for $0 \leq h \leq 1$ and $\tilde{\eta} = (\tilde{\eta}_a, \tilde{\eta}_b)'$, and $\tilde{h}(x, \tilde{\eta}) = \tilde{\eta}_a + \tilde{\eta}_b x - 1$. Then

$$\begin{aligned}
\mathrm{E}[f_{\tilde{v}}(-\tilde{h}(X_2, \tilde{\eta}) \mid \tilde{\eta}, \mathbf{X})\partial_x \tilde{h}(X_2, \tilde{\eta}) \mid \mathbf{X}] &= \mathrm{E}[\tilde{\eta}_b \mid \mathbf{X}] = [P_2(\mathbf{X}) - P_1(\mathbf{X})]/(X_2 - X_1) \\
&= -\mathrm{E}\left[\frac{F_v(-h(X_2, \eta_2) \mid \eta_2, \mathbf{X}) - F_v(-h(X_1, \eta_2) \mid \eta_2, \mathbf{X})}{X_2 - X_1} \mid \mathbf{X}\right] \\
&\neq \mathrm{E}[f_v(-h(X_2, \eta_2) \mid \eta_2, \mathbf{X})\partial_x h(X_2, \eta_2) \mid \mathbf{X}],
\end{aligned}$$

where the last not equal holds when

$$\mathrm{E}\left[\frac{F_v(-h(X_2, \eta_2) \mid \eta_2, \mathbf{X}) - F_v(-h(X_1, \eta_2) \mid \eta_2, \mathbf{X})}{X_2 - X_1} + f_v(-h(X_2, \eta_2) \mid \eta_2, \mathbf{X})\partial_x h(X_2, \eta_2) \mid \mathbf{X}\right] \neq 0.$$

THEOREM 10: *If i)* $P_t(\mathbf{X}) = \mathrm{E}[1 - F_v(-h(X_t, \eta_t) \mid \eta_t, \mathbf{X}) \mid \mathbf{X}]$ *and* ε_t *is time stationary; ii)* $f_v(h \mid \eta_t, \mathbf{X})$ *is bounded; iii)* $h(x, \eta)$ *is continuously differentiable in* x *with bounded derivative; and iv) there does not exist any* $\Delta(\mathbf{X})$ *such that*

$$\mathrm{E}\left[F_v(-h(X_2, \eta_2) \mid \eta_2, \mathbf{X}) - F_v(-h(X_1, \eta_2) \mid \eta_2, \mathbf{X}) \mid \mathbf{X}\right] = \Delta(\mathbf{X})(X_2 - X_1)$$

for $X_1 \neq X_2$, *then* $\mathrm{E}\left[f_v(-h(X_2, \eta_2) \mid \eta_2, \mathbf{X})\partial_x h(X_2, \eta_2) \mid \mathbf{X}\right]$ *is not identified on the set* $X_1 \neq X_2$.

7 Conclusion

Jerry Hausman pioneered the introduction of flexible forms of unobserved heterogeneity in structural economic models for multinomial choice. This paper follows this tradition by considering identification of nonseparable multinomial choice models with unobserved heterogeneity that is unrestricted in both the dimension and its interaction with observed attributes. Some of our results are of local nature. For example, we show that derivatives of choice probabilities identify average utility derivatives only for marginal units that are indifferent between two choices with cross-section data and for units that have time invariant attributes with time stationary panel data. It would be interesting to characterize minimal conditions that permit extending the identification of average utility derivatives to larger populations. We leave this extension to future work.

8 Appendix: Proofs of Theorems

Proof of Theorem 1: The proof is similar to the proof of Lemma 1 of Chernozhukov et al. (2015). Let $F_v(v \mid \eta) = \int_{-\infty}^{v} f_v(u \mid \eta)du$. Under Assumptions 1 and 2,

$$\begin{aligned} P(x) = \Pr(Y = 1 \mid X = x) &= \int \mathbf{1}\{\delta(x, \eta, v) \geq 0\} F_v(dv \mid \eta) F_\eta(d\eta) \\ &= \int \mathbf{1}\{v \geq \delta^{-1}(x, \eta, 0)\} F_v(dv \mid \eta) F_\eta(d\eta) \\ &= 1 - \int F_v(\delta^{-1}(x, \eta, 0) \mid \eta) F_\eta(d\eta). \end{aligned}$$

Differentiating with respect to x

$$\partial_x P(x) = -\int f_v(\delta^{-1}(x, \eta, 0) \mid \eta) \partial_x \delta^{-1}(x, \eta, 0) F_\eta(d\eta),$$

where the conditions of Assumption 2 allow us to differentiate under the integral. Note that by the inverse and implicit function theorems,

$$\partial_x \delta^{-1}(x, \eta, 0) = -\left. \frac{\partial_x \delta(x, \eta, v)}{\partial_v \delta(x, \eta, v)} \right|_{\delta(x,\eta,v)=0}.$$

Also, by a change of variable

$$\left. \frac{f_v(v \mid \eta)}{\partial_v \delta(x, \eta, v)} \right|_{\delta(x,\eta,v)=0} = f_{\delta(x,\eta,v)}(0 \mid \eta),$$

where $f_{\delta(x,\eta,v)}(\cdot \mid \eta)$ is the conditional density of $\delta(x,\eta,v)$ given η. Then substituting in gives

$$\begin{aligned}\partial_x P(x) &= \int f_{\delta(x,\eta,v)}(0 \mid \eta)\partial_x \delta(x,\eta,v)|_{\delta(x,\eta,v)=0} F_\eta(d\eta) \\ &= \mathrm{E}[\partial_x \delta(x,\eta,v) \mid \delta(x,\eta,v) = 0] \cdot f_{\delta(x,\eta,v)}(0) \\ &= \mathrm{E}[\partial_x \delta(x,\varepsilon) \mid \delta(x,\varepsilon) = 0] \cdot f_{\delta(x,\varepsilon)}(0),\end{aligned}$$

since

$$\begin{aligned}\mathrm{E}[\partial_x \delta(x,\eta,v) \mid \delta(x,\eta,v) = 0] &= \int \partial_x \delta(x,\eta,\delta^{-1}(x,\eta,0)) dF_\eta(\eta \mid \delta(x,\eta,v) = 0) \\ &= \int \partial_x \delta(x,\eta,v)|_{\delta(x,\eta,v)=0} \frac{f_{\delta(x,\eta,v)}(0 \mid \eta)}{f_{\delta(x,\eta,v)}(0)} F_\eta(d\eta),\end{aligned}$$

by the Bayes rule. *Q.E.D.*

Proof of Corollary 2: Given in text.

Proof of Corollary 3: Given in text.

Proof of Theorem 4: By iterated expectations,

$$P_j(x) = \mathrm{E}[p_j(u(x,\eta) \mid \eta)] = \int p_j(u(x,\eta) \mid \eta) F_\eta(d\eta).$$

Also by Assumption 3 and the chain rule, $p_j(u(x,\eta) \mid \eta)$ is continuously differentiable in x with bounded derivative

$$\sum_{k=1}^{J} p_{jk}(u(x,\eta) \mid \eta)\partial_x u_k(x,\eta).$$

Interchanging the order of differentiation and integration is then allowed, and the conclusion follows. *Q.E.D.*

Proof of Theorem 5: By iterated expectations and independence of v and x given w

$$P_j(x,w) = \mathrm{E}[p_j(u(x,\eta) \mid \eta, w) \mid w] = \int p_j(u(x,\eta) \mid \eta, w) F_\eta(d\eta \mid w).$$

Also, by $f(v \mid \eta, w)$ continuous in v and bounded and the chain rule, $p_j(u(x,\eta) \mid \eta, w)$ is continuously differentiable in x with bounded derivative

$$\sum_{k=1}^{J} p_{jk}(u(x,\eta) \mid \eta, w)\partial_x u_k(x,\eta).$$

Interchanging the order of differentiation and integration is then allowed, and the conclusion follows. Q.E.D.

Proof of Corollary 6: Given in text.

Proof of Theorem 7: By iterated expectations,

$$P_{jt}(\mathbf{X}) = \mathrm{E}[p_j(u(X_t, \eta_t) \mid \eta_t, \mathbf{X})] = \int p_j(u(X_t, \eta) \mid \eta, \mathbf{X}) F_\eta(d\eta \mid \mathbf{X}),$$

where $F(\eta \mid \mathbf{X})$ denotes the CDF of η_t conditional on $\mathbf{X}$. Also by Assumption 3 and the chain rule $p_j(u(x, \eta) \mid \eta, \mathbf{X})$ is continuously differentiable in x with bounded derivative and

$$\partial_x p_j(u(x, \eta) \mid \eta, \mathbf{X}) = \sum_{k=1}^{J} p_{jk}(u(x, \eta) \mid \eta, \mathbf{X}) \partial_x u_k(x, \eta).$$

It follows by the previous equation that the order of differentiating an integration can be interchanged to obtain

$$\partial_x \mu_j(x \mid \mathbf{X}) = \mathrm{E}\left[\sum_{k=1}^{J} p_{jk}(u(x, \eta_t) \mid \eta_t, \mathbf{X}) \partial_x u_k(x, \eta_t) \mid \mathbf{X} \right].$$

Note that $P_{jt}(\mathbf{X}) = \mu_j(X_t \mid \mathbf{X})$. Then by the chain rule we have

$$\begin{aligned}
\partial_{X_2} \mathrm{E}[Y_{j2} - Y_{j1} \mid \mathbf{X}]\big|_{X_1 = X_2} &= \partial_{X_2} \{\mu_j(X_2 \mid \mathbf{X}) - \mu_j(X_1 \mid \mathbf{X})\}\big|_{X_1 = X_2} \\
&= \partial_x \mu_j(x \mid \mathbf{X})\big|_{x = X_2}\big|_{X_1 = X_2} + \partial_{X_2} \mu_j(x \mid \mathbf{X})\big|_{x = X_2}\big|_{X_1 = X_2} \\
&- \partial_{X_2} \mu_j(x \mid \mathbf{X})\big|_{x = X_1}\big|_{X_1 = X_2} \\
&= \partial_x \mu_j(x \mid \mathbf{X})\big|_{x = X_2}\big|_{X_1 = X_2} \\
&= \mathrm{E}\left[\sum_{k=1}^{J} p_{jk}(u(X_2, \eta_2) \mid \eta_2, \mathbf{X}) \partial_x u_k(X_2, \eta_2) \mid \mathbf{X} \right] \Bigg|_{X_1 = X_2}.
\end{aligned}$$

Interchanging the order of differentiation and integration is then allowed, and the conclusion follows. Q.E.D.

Proof of Corollary 8: Given in text.

Proof of Theorem 9: Given in text.

Proof of Theorem 10: Given in text.

Acknowledgements

We thank the participants at Cambridge-INET and Cemmap Panel Data Workshop for comments and Siyi Luo for capable research assistance. We gratefully acknowledge research support from the NSF. We appreciate the hospitality of the Cowles Foundation where much of the work for this paper was accomplished.

REFERENCES

Abrevaya, J. (2000): "Rank Estimation of a Generalized Fixed-Effects Regression Model," *Journal of Econometrics* 95, 1-23.

Ai, C. (1997): "A Semiparametric Maximum Likelihood Estimator", *Econometrica* 65, 933-963.

Altonji, J., and R. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors", *Econometrica* 73, 1053-1102.

Berry, S., P. Haile (2010): "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," Cowles Foundation Discussion Paper 1718.

Berry, S., J. Levinsohn, A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica* 63, 841-890.

Blundell, R.W. and J.L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," Review of Economic Studies 71, 655–679.

Blomquist, S., A. Kumar, C.-Y. Liang, W.K. Newey (2014): "Individual Heterogeneity, Nonlinear Budget Sets, and Taxable Income," CEMMAP working paper CWP21/14

Burda, M., M. Harding, and J.A. Hausman (2008): "A Bayesian Mixed Logit-Profit Model for Multinomial Choice," *Journal of Econometrics* 147, 232-246.

Burda, M., M. Harding, and J.A. Hausman (2010): "A Poisson Mixture Model of Discrete Choice," *Journal of Econometrics* 166, 184-203.

Chen, S., and X. Wang (2017), "Semiparametric Estimation of a Panel Data Model without Monotonicity or Separability," unpublished manuscript, Hong Kong University of Science and Technology.

Chernozhukov, V., Fernandez-Val, I., Hahn, J., and W. K. Newey (2013), "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), pp. 535–580.

Chernozhukov, V., I. Fernandez-Val, S. Hoderlein, H. Holzman, and W.K. Newey (2015): "Nonparametric Identification in Panels Using Quantiles," *Journal of Econometrics* 188, 378–392.

Chernozhukov, V., I. Fernandez-Val, Y. Luo (2015):"The Sorted Effects Method: Discovering Heterogenous Effects Beyond Their Averages," working paper.

Evdokimov, K. (2010), "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," unpublished manuscript, Princeton University.

Gautier, E. and Y. Kitamura (2013): "Nonparametric Estimation in Random Coefficients Binary Choice Models," *Econometrica* 81, Pages 581–607.

Graham, B.W. and J.L. Powell (2012), "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models," *Econometrica* 80

(5), pp. 2105–2152.

Hausman, J.A., and D. Wise (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica* 46, 403-26.

Hausman, J.A., and W.K. Newey (2016): "Individual Heterogeneity and Average Welfare," *Econometrica* 84, 1225-1248.

Hoderlein, S., and E. Mammen (2007): "Identification of Marginal Effects in Nonseparable Models without Monotonicity," *Econometrica*, 75, 1513 - 1519.

Hoderlein, S. and H. White, (2012), "Nonparametric identi cation in nonseparable panel data models with generalized fixed effects," *Journal of Econometrics* 168, 300-314.

Honore, B.E. (1992): "Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica* 60, 533-565

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics* 58, 71-120.

Imbens, G. and W.K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica* 77, 1481-1512.

Manski, C.F. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica* 55, 357-362.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed) *Frontiers of Econometrics*, Academic Press, 105-142.

McFadden, D.; K. Richter (1991) "Stochastic Rationality and Revealed Stochastic Preference," in J. Chipman, D. McFadden, K. Richter (eds) *Preferences, Uncertainty, and Rationality*, Westview Press, 161-186.

Pakes, A. and J. Porter (2014): "Moment Inequalities for Semi-parametric Multinomial Choice with Fixed Effects," Working paper, Harvard University.

Sasaki, Y. (2015): "What Do Quantile Regressions Identify for General Structural Functions?" *EconometricTheory* 31, 1102-1116.

Shi, X., Shum, M., and W. Song (2017): "Estimating Semi-parametric Panel Multinomial Choice Models using Cyclic Monotonicity," Working paper, University of Wisconsin-Madison.

Stoker, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica* 54, 1461-1482.