Identification and estimation of multinomial choice models with latent special covariates[*]

Nail Kashaev [†]

nkashaev@uwo.ca

First version: November 13, 2018

This version: November 9, 2021

Abstract Identification of multinomial choice models is often established by using special covariates that have full support. This paper shows how these identification results can be extended to a large class of multinomial choice models when all covariates are bounded. I also provide a new $\sqrt{n}$-consistent asymptotically normal estimator of the finite-dimensional parameters of the model.

JEL classification numbers: C50, C57

Keywords: Multinomial choice, random coefficients, special covariate, identification at infinity, bundles

[*] A previous version of this paper was circulated under the title "Identification and estimation of discrete outcome models with latent special covariates." I thank the editor and two anonymous referees for comments and suggestions that have greatly improved the manuscript. I also thank Roy Allen, Victor H. Aguiar, Tim Conley, Nirav Mehta, Salvador Navarro, Joris Pinkse, David Rivers, and Bruno Salcedo for their useful comments and discussions.

[†] Department of Economics, University of Western Ontario.

1. Introduction

This paper studies identification and estimation of random coefficients multinomial choice models with covariates that have bounded support. Often *some* latent variables in these models have full support (i.e. supported on the whole Euclidean space). Under common restrictions on the distribution of these unobservables, I constructively identify it and show how these latent variables can be used to construct special covariates (i.e., artificial observables with full support) to nonparametrically identify the distribution of *all* the other unobservables. Identification of all parts of the structural model is crucial for welfare analysis (e.g., aggregate welfare changes between two choice situations). My identification technique is constructive and leads to an asymptotically normal estimator of the finite-dimensional parameters of the model.

The results of this paper rest on two commonly used assumptions. First, I assume existence of excluded covariates that affect the distribution over choices via a random coefficient. Using variation in these excluded covariates I can identify the distribution of the random coefficient. Second, I assume that the distribution of the random coefficient is sufficiently "rich". "Richness" of the random coefficient distribution is formalized by a notion of bounded completeness.[1] As a result, I show how to identify the distribution over outcomes conditional on the realization of the observed covariates and the latent random coefficient nonparametrically. Since the latent random coefficient often has full support, I can treat it as an observed covariate with full support and apply *any* identification technique that requires existence of such covariates to identify the rest of the model parameters (e.g., the distribution of other latent variables).

I provide two nonnested identification results. The first result does not make any parametric assumptions about the distribution of latent variables. It, however, imposes some restrictions on the support of observables. In particular, I require the support of some covariates to contain zero. It also requires some smoothness of the distribution of the latent variables. To the best of my knowledge, this is the first result in the literature that nonpara-

[1] Completeness of a family of distributions is a well-known concept in the Statistics and Econometrics literature. See, for example, Mattner et al. (1993), Newey and Powell (2003), Chernozhukov and Hansen (2005), Blundell et al. (2007), Chernozhukov et al. (2007), Hu and Schennach (2008), Andrews (2011), Darolles et al. (2011), and d'Haultfoeuille (2011).

metrically identifies the distribution of all latent variables in multinomial choice settings with bounded covariates. The second result uses one of the most popular parameterizations in applied work - a Gaussian distribution of the latent random coefficient. But, in contrast to the first result, it does not require zero in the support of covariates and leaves the distribution of other latent variables completely unrestricted. The second result also leads to an easy to implement asymptotically normal estimator of the finite-dimensional parameters of the model. Similar to Powell et al. (1989), this estimator is $\sqrt{n}$-consistent since it is based on average derivatives of an estimable object.

I contribute to the discrete outcome literature in several respects. I show how existing results that use full-support-excluded covariates with monotonicity restrictions[2] can be directly used in environments with bounded covariates. Formally, I demonstrate that my setting inherits all identifying properties of the setting with a special covariate. I also contribute to the literature on semiparametric models by showing that common parametric restrictions can be used instead of covariates that have full support (e.g., Fox et al., 2012). This paper is also related to the literature on identification of finite-dimensional parameters in discrete outcome models with bounded covariates.[3] The main difference from that literature is that in my framework the distribution of latent variables (e.g., the random intercept) can be nonparametrically identified even if these latent variables have full support, but covariates are bounded.

My approach is complementary to existing methods. Since as an input my framework requires the average structural demand function (i.e., the choice probability function) for one good, my results may be combined with the ones in Berry and Haile (2020) to nonparametricaly identify the distribution of unobserved individual level heterogeneity. Moreover, in situations where the researcher is not sure whether covariates have full support and is willing to impose mild restrictions because of tractability or data limitations, my approach can provide an additional reassurance of identification. Also, the results in this paper provide a more solid econometric foundation to the models with at least one normally distributed random

[2] See, for example, Manski (1985, 1988), Heckman (1990), Matzkin (1992), Ichimura and Thompson (1998), Lewbel (1998, 2000), Tamer (2003), Matzkin (2007), Berry and Haile (2009), Bajari et al. (2010), Gautier and Kitamura (2013), Gautier and Hoderlein (2015), Fox and Gandhi (2016), Dunker et al. (2017), Fox and Lazzati (2017), Fox et al. (2018), Fox (2020), and Kashaev and Salcedo (2021).

[3] E.g., Magnac and Maurin (2007), Chen et al. (2016), Kline (2016), and Lewbel et al. (2021).

coefficient (e.g. Nevo, 2000).

The paper is organized as follows. In Section 2, I describe the setting. Sections 3.1 and 3.2 provide two identification results. I show how my identification results can be extended to bundles model in Section 3.3. In Sections 4 and 5, I propose a new estimator of the finite-dimensional parameters and evaluate its performance in simulations. Section 6 provides an empirical illustration. Section 7 concludes. All proofs can be found in Appendix A. Appendix B provides additional simulation evidence.

2. Multinomial Choice

Consider the following random coefficients model. The agent maximizes (indirect) utility by choosing between J inside goods (e.g., different brands of cereals) and an outside option of no purchase. The choice set is denoted by $Y = \{0, 1, \ldots, J\}$. I normalize the utility from alternative $y = 0$ to 0. The random utility from choosing an alternative $y \neq 0$ is[4]

$$\mathbf{z}_y \left[\beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e}\right] + \boldsymbol{\varepsilon}_y,$$

where $\mathbf{z}_y \in Z_y \subseteq \mathbb{R}$ is a product-specific observed covariate that can be different for different consumers (e.g., fiber content or price); $\mathbf{d} \in D \subseteq \mathbb{R}$ is observed (demographic) individual-specific taste shifter (e.g., age or income); $\mathbf{w} \in W \subseteq \mathbb{R}^{d_w}$ is a vector of all other observable covariates, which may include the rest of product/agent characteristics; $\mathbf{e} \in E \subseteq \mathbb{R}$ is a latent taste shock. The latent random vector $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_y)_{y \in Y \setminus \{0\}}$ captures all other sources of unobserved heterogeneity (e.g., $\boldsymbol{\varepsilon}_y = \boldsymbol{\theta}^\top \mathbf{w}_y + \boldsymbol{\epsilon}_y$ a.s., where $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}_y$ are random coefficients). The observed covariates are $\mathbf{x} = (\mathbf{d}, \mathbf{z}, \mathbf{w})$, where $\mathbf{z} = (\mathbf{z}_y)_{y \in Y \setminus \{0\}}$.

The random coefficient $\beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e}$ represents individual specific heterogeneous tastes associated with the product characteristic $\mathbf{z}_y$ (i.e., the marginal utility from the product

[4] Deterministic vectors are denoted by lower-case regular font Latin letters (e.g., x) and random objects by bold letters (e.g., $\mathbf{x}$). Capital letters are usually used to denote supports of random variables (e.g., $\mathbf{x} \in X$). I denote the support of a conditional distribution of $\mathbf{x}$ conditional on $\mathbf{z} = z$ by X_z. The cumulative distribution function (c.d.f.) and the probability density function (p.d.f.) of $\mathbf{x}$ are denoted by $F_{\mathbf{x}}$ and $f_{\mathbf{x}}$. $F_{\mathbf{x}|\mathbf{z}}$ ($f_{\mathbf{x}|\mathbf{z}}$) denotes the c.d.f. (p.d.f.) of $\mathbf{x}$ conditional on $\mathbf{z} = z$.

characteristic $\mathbf{z}_y$). This specification of random coefficients is common in applied work (see, for instance, Berry et al., 1995, Nevo, 2000, 2001, Berry et al., 2004). The functions $\beta_0, \beta_1 : W \to \mathbb{R}$ are unknown to the researcher and $\beta_1(w) \neq 0$ for all $w \in W$. I assume that $\mathbf{d}$ (and $\beta_1(\mathbf{w})$) is scalar without loss of generality since if $\mathbf{d}$ is a vector, then all components of it but one can be absorbed by w. Similarly to the existing treatment of random coefficients model, I assume that the random coefficients in front of $\mathbf{z}_y$ are the same for each alternative y. However, I do not impose sign restrictions on $[\beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e}]$.[5]

I start by stating two assumptions that will be used throughout the paper. The first one is a data requirement, the second one is a shape constraint on the distribution of latent variables.

Assumption 1 (Data) The analyst can identify $p_0(x) = \Pr(\mathbf{y} = 0|\mathbf{x} = x)$ for all $x \in X$

Assumption 1 implies that I only need to observe whether a consumer bought a product or not without knowing the identity of the product (see also, for instance, Thompson, 1989, Lewbel, 2000, Fox et al., 2012).[6] If the information on the identity of the purchases is also available, then (i) this information may improve the efficiency of an estimator; and (ii) help to satisfy the assumptions needed for identification (e.g., in my empirical illustration, I use one product to identify the sign of β_1 and I use another one to estimate it).

Assumption 2 (Exclusion Restrictions) For all $w \in W$

(i) $\boldsymbol{\varepsilon}$ is conditionally independent of $(\mathbf{e}, \mathbf{d}, \mathbf{z})$ conditional on $\mathbf{w} = w$;

(ii) $\mathbf{e}$ is conditionally independent of $(\mathbf{d}, \mathbf{z})$ conditional on $\mathbf{w} = w$.

Assumption 2 is an exclusion restriction that requires latent shocks $\mathbf{e}$ and $\boldsymbol{\varepsilon}$ to be independent of each other (condition (i)) and independent of excluded covariates $(\mathbf{d}, \mathbf{z})$ (condition (ii)) after conditioning on $\mathbf{w}$. Assumption 2 allows any form of dependence between $(\boldsymbol{\varepsilon}, \mathbf{e})$ and nonexcluded covariates $\mathbf{w}$. That is, $\boldsymbol{\varepsilon}$ may contain latent product characteristics (e.g.,

[5] Since $\Pr(\beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e} > 0|\mathbf{x} = x) = 1 - F_{\mathbf{e}|\mathbf{x}}(-\beta_0(w) - \beta_1(w)d|x)$ and there are no restrictions on $\beta_0(\cdot)$, the random coefficient $[\beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e}]$ can be positive (negative) with probability that is arbitrarily close to 1 if the support of $\mathbf{e}$ conditional on $\mathbf{x} = x$ is unbounded.

[6] The outcome $y = 0$ can be replaced by any outcome. In this case, one will just need to renormalize the utility from that outcome to zero.

unobserved quality) that can be correlated with nonexcluded covariates (e.g., market-product identifier).[7] In general, since I only require the identification of the structural demand function p_0, one can use the results in Berry and Haile (2020) to identify p_0 and treat market-product level unobservables as a part of $\mathbf{w}$.

Next, I provide two nonnested sets of conditions that allow for identification of β_0, β_1, and the distribution of $\mathbf{e}$ and $\boldsymbol{\varepsilon}$. In Section 3.1, I impose no parametric assumptions on latent $\mathbf{e}$ and $\boldsymbol{\varepsilon}$ but assume some smoothness on the c.d.f. of $\boldsymbol{\varepsilon}$ and restrict the support of covariates. In Section 3.2, I identify the model when $\mathbf{e}$ is normally distributed, without any additional restrictions on the distribution of $\boldsymbol{\varepsilon}$ and with minimal support restrictions on covariates.

3. Identification

3.1. Nonparametric Identification

Assumption 3 For all $w \in W$

(i) Conditional on $\mathbf{w} = w$, $\mathbf{e}$ has mean zero and variance one;

(ii) $F_{\boldsymbol{\varepsilon}|\mathbf{w}}(\cdot|w)$ has bounded partial derivatives up to order κ for some $\bar{y}$ and $\partial^l_{\varepsilon^l_{\bar{y}}} F_{\boldsymbol{\varepsilon}|\mathbf{w}}(\cdot|w)|_{\varepsilon=0} \neq 0$ for all $l \leq \kappa$;

(iii) There exists d^* such that the support of $(\mathbf{d}, \mathbf{z})$ conditional on $\mathbf{w} = w$ contains $(d^*, 0)$ with an open neighborhood.

Assumption 3(i) is a scale and location normalization. It restricts $\mathbf{e}$ conditional on $\mathbf{w} = w$ to have a finite expectation and a nonzero variance for all w. Assumption 3(ii) requires the conditional distribution $F_{\boldsymbol{\varepsilon}|\mathbf{w}}$ to be sufficiently smooth in one component of ε in the neighborhood of zero and have different from zero higher order partial derivatives. Since

[7] Since, for identification and estimation, I require the average structural function p_0, some forms of endogeneity (i.e., correlation between $\mathbf{x}$ and $\boldsymbol{\varepsilon}$) can be addressed using suitable instruments and control function residuals as in Blundell and Powell (2004) (see also Berry, 1994, Berry et al., 1995, Berry and Haile, 2014 for identification of structural demand function using aggregate data and instruments). I leave the detailed analysis of this case for future research.

$\mathbb{E}[\boldsymbol{\varepsilon}]$ is not assumed to be zero, if, for instance, $\boldsymbol{\varepsilon}$ is multivariate normal with componentwise nonzero mean, then Assumption 3(ii) is automatically satisfied. It is also generically satisfied when at least one component of $\boldsymbol{\varepsilon}$ is independent of the others and has a type I extreme value distribution (Fox et al., 2012). However, Assumption 3(ii) rules out cases when $\boldsymbol{\varepsilon}$ is a constant. Another example of violation of Assumption 3(ii) is when κ is infinite and $F_{\boldsymbol{\varepsilon}|\mathbf{w}}$ is a polynomial function of any finite degree. (In Section 3.2, I provide an alternative result that does not restrict $F_{\boldsymbol{\varepsilon}|\mathbf{w}}$.) Assumption 3(iii) requires the support of $\mathbf{z}$ to contain zero with some open neighborhood. Assumptions similar to Assumptions 3(ii)-(iii) are common in the literature on identification of random coefficients models (e.g., Assumptions 8 and 10 in Fox et al., 2012 and Assumption 4 in Allen and Rehbeck, 2020).

Proposition 3.1 *If Assumptions 1- 3 hold, then $\beta_0(w)$, $\beta_1(w)$, and $\mathbb{E}\left[\mathbf{e}^l|\mathbf{w}=w\right]$, $0 \leq l \leq \kappa$, are identified for all $w \in W$.*

Identification of $\kappa \leq \infty$ moments of the conditional distribution of $\mathbf{e}$ conditional on $\mathbf{w}$ is often sufficient for nonparametric identification of it. For example, Assumption 7 in Fox et al. (2012) uses the Carleman condition.[8] Thus, under minimal restrictions, I can nonparametically identify the conditional c.d.f. $F_{\mathbf{v}|\mathbf{x}}$, where $\mathbf{v} = \beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e}$.

To establish the next identification result I need the following definition.

Definition 1 (Bounded completeness) The family of distributions $\left\{F_{\mathbf{v}|\mathbf{x}}(\cdot|x), x \in X'\right\}$ is boundedly complete if

$$\forall x \in X',\ \int_V g(t)dF_{\mathbf{v}|\mathbf{x}}(t|x) = 0 \implies g(\mathbf{v}) = 0 \text{ a.s.},$$

for any bounded function g.

Completeness assumptions have been widely used in econometric analysis. Completeness is typically imposed on the distribution of observables (e.g., Newey and Powell, 2003). However, many commonly used parametric restrictions on the distribution of unobservables imply bounded completeness. For instance, it is satisfied for normal distributions and the Gumbel

[8] For more detailed discussion of the problem of identification of the distribution from its moments see, for instance, Kleiber and Stoyanov (2013) and references therein.

distribution.[9]

Combining bounded completeness with the identified distribution of the index $\mathbf{v}$, I have the following result.

Proposition 3.2 *If $F_{\mathbf{v}|\mathbf{x}}$ is identified and $\left\{F_{\mathbf{v}|\mathbf{x}}(\cdot|(d,z,w)), d \in D_{(z,w)}\right\}$ is boundedly complete for all (z,w) in the support, then the above model inherits all identifying properties of the random coefficients model with utilities $\mathbb{1}(y \neq 0)(\mathbf{r}_y + \boldsymbol{\varepsilon}_y)$. The vector $\mathbf{r} = (\mathbf{r}_y)_{y \in Y \setminus \{0\}}$ is an observed covariate conditionally independent of $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_y)_{y \in Y \setminus \{0\}}$ conditional on $\mathbf{w} = w$ with the conditional support $R_w = \left\{r \in \mathbb{R}^J \,:\, r = vz,\ z \in Z_w, v \in V_w\right\}$, where V_w is the support of $\mathbf{v}$ conditional on $\mathbf{w} = w$. In particular, $F_{\boldsymbol{\varepsilon}|\mathbf{w}}$ is identified over R_w.*

The proof of Proposition 3.2 is similar to the proof of Theorem 11 in Fox et al. (2012). The main difference is that, instead of parametric restrictions, Proposition 3.2 uses the interaction between $\mathbf{d}$ and $\mathbf{z}$.

Proposition 3.2 implies that the original random coefficient model can be represented in the "special-covariate-with-full-support" framework without assuming existence of such covariates. Moreover, if the set of directions that $z/\left\|z\right\|$ can cover is sufficiently rich and the support of $\mathbf{e}$ conditional on $\mathbf{w} = w$ is $\mathbb{R}$, then $R_w = \mathbb{R}^J$ and all the identification results that require existence of special covariates with full support (e.g., Lewbel, 2000, Berry and Haile, 2009, Gautier and Hoderlein, 2015, Fox and Gandhi, 2016, and Fox, 2020) can be applied.

Combining the results in Propositions 3.1 and 3.2 with Theorem 1 in Fox (2020), I can establish the following result.

Corollary 3.3 *For all $y \neq 0$, let $\boldsymbol{\varepsilon}_y = \boldsymbol{\theta}^{\mathsf{T}}\mathbf{w}_y + \boldsymbol{\zeta}_y$, where $\boldsymbol{\theta}$ and $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_y)_{y \in Y \setminus \{0\}}$ are random coefficients, and $\mathbf{w}_y$ is the vector of product-y-specific covariates. Suppose*

(i) The assumptions of Propositions 3.1 and 3.2 hold;

(ii) $R_w = \mathbb{R}^J$ for all $w \in W$;

(iii) $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ and $\mathbf{w} = (\mathbf{w}_y)_{y \in Y \setminus \{0\}}$ are independent;

(iv) The support of $\mathbf{w}$ contains an open ball of dimensionality of $\mathbf{w}$;

[9] For testability of the completeness assumptions see Canay et al. (2013).

(v) $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ *has finite absolute moments and its distribution is uniquely determined by its moments;*

then β_0*,* β_1*, and the distribution of* $(\mathbf{e}_1, \boldsymbol{\theta}, \boldsymbol{\epsilon})$ *are identified.*

To the best of my knowledge, Corollary 3.3 is the first result that establishes nonparametric identification of the whole distribution of the random coefficients in the multinomial choice environments without assuming the existence of special covariates. Fox et al. (2012), Allen and Rehbeck (2020), and Lewbel et al. (2021) also allow for bounded covariates. However, they either do not fully identify the distribution of the random intercept $\boldsymbol{\varepsilon}$ (Allen and Rehbeck, 2020, Lewbel et al., 2021) or impose parametric restrictions on it (Fox et al., 2012).

3.2. Normal Taste Shock

Assumption 4 For all $w \in W$

(i) Conditional on $\mathbf{w} = w$, $\mathbf{e}$ is a standard normal random variable;

(ii) there exists $(d^*, z^{*\intercal})$ in the interior of the support of $(\mathbf{d}, \mathbf{z})$ conditional on $\mathbf{w} = w$ such that $z_y^* > 0$ for all $y \in Y$;

(iii) there exists $(d^{**}, z^{**\intercal})$ in the interior of the support of $(\mathbf{d}, \mathbf{z})$ conditional on $\mathbf{w} = w$ such that $p_0((d, z^{**}, w))$ is neither an exponential nor an affine function of d on some open set.

Assumption 4(i) requires $\mathbf{e}$ to be normally distributed with nonzero variance. With nonzero variance, the assumption that $\mathbb{E}\left[\mathbf{e}^2\right] = 1$ is just a scale normalization. The assumption is common in applied work (e.g., Nevo, 2000, 2001) and allows me to relax Assumptions 3(ii)-(iii). Assumption 4(ii) is only needed for identification of the sign of $\beta_1(w)$. Assumption 4(iii) means that if I fix all covariates but the one that shifts the random coefficient, then the probability of the default conditional on covariates is neither an affine nor an exponential function of this nonfixed covariate. Assumption 4(iii) is not very restrictive since it rules out only some exponential and linear probability models. Moreover, it is testable.

Proposition 3.4 *If Assumptions 1, 2, and 4 hold, then*

(i) $\beta_0(w)$ and $\beta_1(w)$ are identified for all $w \in W$;

(ii) The conditions of Proposition 3.2 are satisfied.

The proof of the identification of β_0 and β_1 uses the multiplicative structure of d and z, and properties of the standard normal p.d.f. Informally, note that

$$\beta_0(\mathbf{w})\mathbf{z} + \beta_1(\mathbf{w})\mathbf{dz} + \mathbf{ez}.$$

Since d and z can be moved independently, I can use variation in d while keeping dz by varying z to identify $\beta_0(w)$. Then, by varying z, I can identify $\beta_1(w)$. Proposition 3.4(ii) follows from $\beta_0(w)$ and $\beta_1(w)$ being identified and $\mathbf{e}$ being standard normal (i.e, $\beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e}$ conditional on $\mathbf{x} = x$ generates a boundedly complete family of distributions).

Note that the only restriction on $\boldsymbol{\varepsilon}$ needed for Proposition 3.4 is the conditional independence assumption (Assumption 2). The random intercept $\boldsymbol{\varepsilon}$ is allowed be continuously or discretely distributed (e.g., it may be a constant). Hence, I can extend Theorem 2 in Fox and Gandhi (2016) to environments with bounded covariates.

Corollary 3.5 *For all $y \neq 0$ let $\boldsymbol{\varepsilon}_y = \boldsymbol{\theta}_y(\mathbf{w})$, where $\boldsymbol{\theta}_y$ is a random function such that its realization θ_y is a map from W to $\mathbb{R}$. Suppose*

(i) Assumptions of Proposition 3.4 hold;

(ii) $R_w = \mathbb{R}^J$ for all $w \in W$;

(iii) $\boldsymbol{\theta} = (\boldsymbol{\theta}_y)_{y\neq 0}$ and $\mathbf{w}$ are independent;

(iv) The support of $\boldsymbol{\theta}$, Θ, satisfies Assumption 4 in Fox and Gandhi (2016);

then β_0, β_1, and the distribution of $\boldsymbol{\theta}$ are identified.

3.3. Bundles

Note that since I do not assume independence among $\boldsymbol{\varepsilon}_y$ across y, the multinomial choice model I study covers some bundles models (Gentzkow, 2007, Dunker et al., 2017, Fox and

Lazzati, 2017). In particular, assume that there are $\tilde{J}$ goods and the agent can purchase any bundle consisting of these goods. The vector $\tilde{y}$ describes the purchasing decision of the agent. That is, $\tilde{y} \in \tilde{Y} = \{0,1\}^{\tilde{J}}$. For instance, $\tilde{y} = (0,1,0,1,0,\dots,0)$ corresponds to the case when the agent purchased a bundle of goods 2 and 4. The random utility from choosing bundle $\tilde{y} \neq 0$ is of the form

$$(\beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e}) \sum_{j=1}^{\tilde{J}} \tilde{y}_j \tilde{\mathbf{z}}_j + \boldsymbol{\varepsilon}_{\tilde{y}},$$

and the utility from buying nothing is zero. I can rewrite the above utilities from bundles as as the utilities form the multinomial choice problem since there are finitely ($2^{\tilde{J}}$) possible bundles. Indeed, I can enumerate them all with $y = 0$ corresponding to $\tilde{y} = 0 \in \mathbb{R}^{\tilde{J}}$ (i.e., $Y = \{0,1,2,\dots,2^{\tilde{J}}\}$) and define $z_y = \sum_{j=1}^{\tilde{J}} \tilde{y}_j \tilde{z}_j$. As a result, I can extend the conclusions of Theorem 1 in Fox and Lazzati (2017) to environments with bounded covariates

Corollary 3.6 *Let $J = 2$ and*

$$\begin{aligned} \boldsymbol{\varepsilon}_{(1,0)} &= \theta_1(\mathbf{w}) + \boldsymbol{\epsilon}_1, \qquad \boldsymbol{\varepsilon}_{(0,1)} = \theta_2(\mathbf{w}) + \boldsymbol{\epsilon}_2, \\ \boldsymbol{\varepsilon}_{(1,1)} &= \boldsymbol{\varepsilon}_{(1,0)} + \boldsymbol{\varepsilon}_{(0,1)} + \boldsymbol{\xi}\theta_3(\mathbf{w}), \end{aligned}$$

where $\theta_i(\cdot)$, $i = 1,2,3$, are some unknown functions, and $(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \boldsymbol{\xi}) \in \mathbb{R}^2 \times \mathbb{R}_+$. Suppose

(i) Assumptions of Propositions 3.1 and 3.2 or Proposition 3.4 hold;

(ii) $R_w = \mathbb{R}$ for all w;

(iii) $(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2)|\mathbf{w} = w$ has an everywhere positive Lebesgue density on its support for all $w \in W$;

(iv) $\mathbb{E}[\boldsymbol{\epsilon}_i|\mathbf{w} = w] = 0$ and $\mathbb{E}[\boldsymbol{\xi}|\mathbf{w} = w] = 1$ for all $w \in W$ and $i = 1,2$,

then $\theta_i(\cdot)$, $i = 1,2,3$, and the c.d.fs $F_{\boldsymbol{\epsilon}_i|\mathbf{w}}$, $i = 1,2$, and $F_{\boldsymbol{\xi}|\mathbf{w}}$ are identified.

4. Estimation of β

Proposition 3.4 constructively identifies β_0 and β_1. In this section, I use it to estimate these parameters. That is, I focus on the multinomial choice model with random coefficients with normally distributed $\mathbf{e}$.[10] Moreover, to simplify the exposition, I assume that there are no nonexcluded covariates $\mathbf{w}$ (i.e., $\beta_0(\cdot)$ and $\beta_1(\cdot)$ are constant functions). Note that, even though β_0 and β_1 are finite-dimensional parameters and the distribution of $\mathbf{e}$ is assumed to be known, the model is still semiparametric since the distribution of $\boldsymbol{\varepsilon}$ is not parametric.

The first ingredient of the estimator is a nonparametric estimator of $p_0(\cdot) = \Pr(\mathbf{y} = 0|\mathbf{x} = \cdot)$, $\hat{p}_0(\cdot)$. Any consistent and smooth enough estimator $\hat{p}_0$ will deliver a consistent estimator of $\beta = (\beta_1, \beta_0)$.[11] For concreteness, I work with the series estimator based on products of powers of components of $x = (d, z)$ (polynomial regressions). That is, given a sample of independent identically distributed (i.i.d.) observations on covariates and a binary random variable that indicates whether the product was purchased or not $\left\{\mathbb{1}\left(\mathbf{y}^{(i)} = 0\right), \mathbf{x}^{(i)}\right\}_{i=1}^{n}$, define

$$\hat{p}_0(x) = \psi^K(x)^\intercal \left(\Psi^\intercal \Psi\right)^- \sum_{i=1}^{n} \psi^K\left(\mathbf{x}^{(i)}\right) \mathbb{1}\left(\mathbf{y}^{(i)} = 0\right),$$

where $\psi^K(\cdot)$ is a vector of orthonormal basis functions based on products of powers of components of x, $\Psi = \left(\psi^K\left(\mathbf{x}^{(1)}\right), \psi^K\left(\mathbf{x}^{(2)}\right), \ldots, \psi^K\left(\mathbf{x}^{(n)}\right)\right)^\intercal$, and $\left(\Psi^\intercal\Psi\right)^-$ is the Moore-Penrose generalized inverse. I assume that the sum of powers of components of x in ψ^K is monotonically increasing in K.

The sign of β_1 can be trivially estimated from $\hat{p}_0$ since

$$\operatorname{sign}(\beta_1) = \operatorname{sign}\left(p_0((d', z)) - p_0((d, z))\right) \operatorname{sign}(z_{y^*}) \operatorname{sign}(d' - d)$$

if $z \geq 0$ or $z \leq 0$ with $z_{y^*} \neq 0$. Hence, for simplicity I assume that $\beta_1 > 0$.

The identification result in Proposition 3.4 is constructive and provides a closed form expression for β as a functional of p_0 (see Appendix A.3). Given the nonparametric power

[10] Proposition 3.1 also provides a constructive identification for β_0 and β_1. However, Assumption 3(iii) fails to hold in my illustrative application presented in Section 6. Additionally, Proposition 3.1 uses limits of derivatives of identifiable functions at a single point, thus, most likely, leading to a consistent estimator with nonparametric rate of convergence.

[11] The normality of $\mathbf{e}$ implies that p_0 has continuous derivatives of any order. See Appendix A.3 for details.

series estimator $\hat{p}_0$, the plug-in estimator of β is

$$\hat{\beta}_1 = \sqrt{\frac{\sum_{i=1}^{n} \hat{p}_{111}\left(\mathbf{x}^{(i)}\right)\hat{p}_1\left(\mathbf{x}^{(i)}\right) - \hat{p}_{11}\left(\mathbf{x}^{(i)}\right)^2}{\sum_{i=1}^{n} \hat{p}_{12}\left(\mathbf{x}^{(i)}\right)\hat{p}_1\left(\mathbf{x}^{(i)}\right) - \hat{p}_2\left(\mathbf{x}^{(i)}\right)\hat{p}_{11}\left(\mathbf{x}^{(i)}\right) - \hat{p}_1\left(\mathbf{x}^{(i)}\right)^2}},$$
$$\hat{\beta}_0 = \hat{\beta}_1 \frac{\sum_{i=1}^{n} \hat{p}_2\left(\mathbf{x}^{(i)}\right) - \mathbf{d}^{(i)}\hat{p}_1\left(\mathbf{x}^{(i)}\right)}{\sum_{i=1}^{n} \hat{p}_1\left(\mathbf{x}^{(i)}\right)} - \frac{1}{\hat{\beta}_1}\frac{\sum_{i=1}^{n} \hat{p}_{11}\left(\mathbf{x}^{(i)}\right)}{\sum_{i=1}^{n} \hat{p}_1\left(\mathbf{x}^{(i)}\right)},$$

where

$$\hat{p}_1(x) = \partial_d \hat{p}_0(x), \quad \hat{p}_{11}(x) = \partial_{d^2}^2 \hat{p}_0(x), \quad \hat{p}_{111}(x) = \partial_{d^3}^3 \hat{p}_0(x),$$
$$\hat{p}_2(x) = \sum_{y=1}^{J} z_y \partial_{z_y} \hat{p}_0(x), \quad \hat{p}_{12}(x) = \partial_d \hat{p}_2(x).$$

Note that $\hat{\beta}$ is essentially a nonlinear function of sample averages of different derivatives of estimated $\hat{p}_0$. Following Newey (1994, 1997), to achieve $\sqrt{n}$-consistency and asymptotic normality of the proposed estimator, I will have to establish existence of the Reisz representer of a particular directional derivative. Let

$$\bar{v}_1(x) = -\left[4p_{1111}(x)f_{\mathbf{x}}(x) + 8p_{111}(x)\partial_d f_{\mathbf{x}}(x) + 5p_{11}(x)\partial_{d^2}^2 f_{\mathbf{x}}(x) + p_1(x)\partial_{d^3}^3 f_{\mathbf{x}}(x)\right]/f_{\mathbf{x}}(x),$$
$$\bar{v}_2(x) = \left[\beta_1\{(1-J)f_{\mathbf{x}}(x) + d\partial_d f_{\mathbf{x}}(x) - \sum_y z_y \partial_{z_y} f_{\mathbf{x}}(x)\} - \partial_{d^2}^2 f_{\mathbf{x}}(x)\right]/f_{\mathbf{x}}(x),$$
$$\bar{v}(x) = (\bar{v}_1(x), \bar{v}_2(x)),$$

where $f_{\mathbf{x}}$ is the p.d.f. of $\mathbf{x}$, and p_1, p_{11}, p_{111}, and p_{1111} are first, second, third, and forth derivatives of p_0 with respect to d, respectively.

Assumption 5 (i) The support of $\mathbf{x}$, X, is a Cartesian product of compact connected nonsingleton intervals in $\mathbb{R}$.

(ii) $f_{\mathbf{x}}$ is bounded away from zero on the interior of X;

(iii) $f_{\mathbf{x}}$, $\partial_d f_{\mathbf{x}}$, $\partial_{z_y} f_{\mathbf{x}}$, and $\partial_{d^2}^2 f_{\mathbf{x}}$ equal to zero at the boundary of X for all y;

(iv) $\mathbb{E}\left[\bar{v}(\mathbf{x})\bar{v}(\mathbf{x})^{\mathsf{T}}\right]$ is finite and nonsingular.

Assumptions 5(i)-(ii) are standard in the literature on nonparametric estimation of conditional expectations. Similarly to the average derivative estimator of Powell et al. (1989), to achieve $\sqrt{n}$-consistency the estimator I need to impose restrictions on the behavior of $f_{\mathbf{x}}$ on the boundary of its support. Since Powell et al. (1989) work with the first derivative they only require $f_{\mathbf{x}}$ to vanish on the boundary. My estimator involves derivatives up to order 3, thus, leading to Assumption 5(iii). Assumption 5(iv) is the mean-square continuity condition that requires the variance of the score function of $\mathbf{x}$ (i.e $\log f_{\mathbf{x}}$) and derivatives of it to be finite.

The following proposition establishes asymptotic normality of my estimator and is based on Theorem 6 in Newey (1997). Denote

$$G = \begin{pmatrix} 2\beta_1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbb{E}\left[\, p_{12}(\mathbf{x})p_1(\mathbf{x}) - p_2(\mathbf{x})p_{11}(\mathbf{x}) - p_1(\mathbf{x})^2 \,\right] & 0 \\ 0 & \beta_1 \mathbb{E}\left[\, p_1(\mathbf{x}) \,\right] \end{pmatrix}^{-1},$$

Proposition 4.1 *If (i) $\left\{\mathbb{1}\left(\mathbf{y}^{(i)} = 0\right), \mathbf{x}^{(i)}\right\}_{i=1}^{n}$ are i.i.d.; (ii) Assumptions 2, 4 and 5 are satisfied, and Assumption 4(iii) is satisfied for all $x^{**} = (d^{**}, z^{**}) \in X$; (iii) $K^6/n \to_{n\to\infty} 0$, then*

$$\sqrt{n}(\hat{\beta} - \beta) \to_d \mathrm{N}(0, V),$$

where $V = G\mathbb{E}\left[\bar{v}(\mathbf{x})\bar{v}(\mathbf{x})^{\mathsf{T}} p_0(\mathbf{x})(1 - p_0(\mathbf{x}))\right] G^{\mathsf{T}}$.

In the proof of Proposition 4.1, I also provide a consistent estimator of the asymptotic variance matrix V that is based on the estimator proposed in Newey (1997).

I conclude this section by noting that after β is estimated, one can construct a sieve maximum-likelihood estimator of F_ε since

$$\Pr(\mathbf{y} = 0|\mathbf{x} = x) = \int_{\mathbb{R}} F_\varepsilon(tz_1, tz_2, \ldots, tz_J)\phi\left(t + \beta_0 + \beta_1 d\right) dt$$

where $\phi(\cdot)$ is the standard normal p.d.f. Thus, one can find the maximizer of

$$\max_{F \in \mathcal{F}_n} \sum_{i=1}^{n} \mathbb{1}\left(\mathbf{y}^{(i)} = 0\right) \log\left(\int_{\mathbb{R}} F(tz_1, tz_2, \ldots, tz_J)\phi\left(t + \hat{\beta}_0 + \hat{\beta}_1 d\right) dt\right) +$$

$$\mathbb{1}\left(\mathbf{y}^{(i)} \neq 0\right) \log \left(1 - \int_{\mathbb{R}} F(tz_1, tz_2, \ldots, tz_J)\phi\left(t + \hat{\beta}_0 + \hat{\beta}_1 d\right) dt\right),$$

where $\{\mathcal{F}_n\}_{n=1}^{\infty}$ is a sequence of sieve spaces for F_ε. Inference on known functionals of β and F_ε (e.g., counterfactuals) can be done using likelihood-ratio type statistic (see, for instance, Shen and Shi, 2005, Chen and Liao, 2014).[12]

5. Monte-Carlo Simulations

In this section, I assess the performance of my estimator in finite samples. I consider the binary choice model:

$$\mathbf{y} = \mathbb{1}\left((\beta_0 + \beta_1 \mathbf{d} + \mathbf{e})\mathbf{z} + \beta_3 + \boldsymbol{\varepsilon} \geq 0 \right),$$

where $\beta_0 = -0.5$, $\beta_1 = 1$, and $\mathbf{e}$ is a standard normal random variable. The random intercept $\beta_3 + \boldsymbol{\varepsilon}$ is independent from $\mathbf{x}$ and $\mathbf{e}$ with mean $\beta_3 = 0.5$. The observed covariates $\mathbf{x} = (\mathbf{d}, \mathbf{z})$ are distributed according to a monotone transformation of a bivariate normal distribution: $\mathbf{x} = 5(\arctan(\tilde{\mathbf{x}})/\pi + 0.5)$, where $\tilde{\mathbf{x}}$ is a mean-zero normal random vector such that each component of it has variance 1 and the correlation between components is 0.1. Note that $\mathbf{x}$ has bounded support.

I consider several data generating processes (DGPs). The first one (DGP-0) is when $\boldsymbol{\varepsilon}$ is a standard normal random variable. The next five DGPs correspond to $\boldsymbol{\varepsilon}$ being an equally weighted mixture of three unit-variance normal distributions with mean $-t$, 0, and t for $t \in \{1, 2, 3, 4, 5\}$ (DGP-t). For every t the distribution of $\boldsymbol{\varepsilon}$ is symmetric. However, the variance is growing with t and the distribution changes from a unimodal distribution to a distribution with three modes. Finally, DGP-L corresponds to the case with logistically distributed $\boldsymbol{\varepsilon}$.

Each experiment is conducted 1000 times for every DGP for 3 sample sizes $n \in \{10^3, 5 \cdot 10^3, 10^4\}$. I use a tensor product of cubic polynomials in estimation of the conditional probability p_0.[13] The results for the mean deviation (bias) of the estimator of β_1 are presented in

[12] Both β and F_ε can be estimated in one step by the sieve maximum-likelihood estimator. In this case, however, the estimator of β may not be $\sqrt{n}$ consistent.

[13] The results are qualitatively the same for higher order polynomials.

Table 1. As expected, the bias decreases with the sample size.[14] However, there is not much variation across DGPs.[15]

Table 1 – Bias

Sample Size	DGP-0	DGP-1	DGP-2	DGP-3	DGP-4	DGP-5	DGP-L
1000	1.08	1.10	1.18	1.46	1.51	1.50	1.12
5000	0.36	0.54	0.89	1.05	1.25	1.23	0.62
10000	0.17	0.26	0.57	0.84	1.09	1.17	0.38

6. Illustrative Empirical Application

To illustrate the empirical importance of the relaxation of the parametric assumptions about the distribution of F_ε and the proposed estimation procedure, I analyze margarine purchasing decisions of households from Springfield, MO, USA, using the multinomial choice model with normally distributed $\mathbf{e}$. I find substantial differences between estimates obtained by employing my semiparametric estimator and a fully parametric multinomial-logit-type estimator.

Data

The original dataset, constructed by Allenby and Rossi (1991), is a panel of 9196 purchases of 10 brands of stick and tube margarine by 517 households from Springfield, MO, USA, extracted from an ERIM (A.C. Nielsen) scanner dataset. The dataset contains information on the shelf prices of each brand that is constructed using the actual price paid and the value of any redeemed coupon. The household demographics contain information on the household income.[16] Benoit et al. (2016) focused on 5 brands instead of 10 and transformed this dataset to a cross-section with 242 households. In particular, every observation contains

[14] The mean absolute deviation of the estimator also decreases with the sample size. See, Appendix B for further details.

[15] For comparison of my estimator with two alternative potentially misspecified parametric estimators, see Appendix B.

[16] See Allenby and Rossi (1991) for specific details of the dataset construction.

only information on the household annual income, which I use as the agent-specific covariate d, agent choices (y), and product-specific prices p_y.[17] There are 5 brands: Generic ($y = 0$), Blue Bonnet ($y = 1$), House Brand ($y = 2$), Shed Spread ($y = 3$), and Fleischmann's ($y = 4$).

Income varies from 2.5k to 130k, with the median and average income being 26.75k and 22.5k, respectively. Table 2 summarizes the share and price information for different products. There is a variation in prices across brands with Generic being on average the cheapest and Fleischmann's being the most expensive. At the same time, Fleischmann's is the least demanded product.

Table 2 – Summary Statistics for Products

Brand	Share	Average Price	Median Price	Max Price	Min Price
Generic	0.17	0.37	0.36	0.33	0.53
Blue Bonnet	0.30	0.58	0.61	0.19	0.76
House Brand	0.19	0.51	0.57	0.19	0.58
Shed Spread	0.21	0.83	0.85	0.50	0.98
Fleischmann's	0.12	1.04	1.08	0.99	1.13

Utility

I follow Nevo (2000, 2001) and model the utility from purchasing brand $y \in \{0, 1, 2, 3, 4\}$ as

$$\boldsymbol{\delta}\mathbf{d} + (\beta_0 + \beta_1\mathbf{d} + \mathbf{e})\mathbf{p}_y + \tilde{\boldsymbol{\varepsilon}}_y.$$

The random coefficient $\boldsymbol{\delta}$ captures the direct marginal effect of income on utility from consumption of margarine (i.e., it is the same for all brands). The coefficient $\beta_0 + \beta_1\mathbf{d}$ can be thought of as the average marginal utility with respect to price. It captures the sensitivity of agents with respect to prices and is expected to be negative. Agents with different incomes may react differently to price changes. Note that no assumptions are made about $\tilde{\boldsymbol{\varepsilon}}_y$ (e.g., it is not assumed that it is has zero mean).[18] This utility specification correspond to the "preference shifter" specification in Griffith et al. (2018). There is no information about those who did not purchase any margarine products, thus, I analyze the choices of those who already

[17] Income and prices are measured in thousands of US dollars and US dollars, respectively.

[18] Estimation using $\log(\mathbf{p}_y)$ instead of $\mathbf{p}_y$ gives qualitatively similar results.

decided to purchase a margarine product. If I treat the utility from consuming Generic brand as the baseline utility and subtract it from all utilities, the normalized utility from purchasing different brands for $y = 1, 2, 3, 4$ is

$$(\beta_0 + \beta_1 \mathbf{d} + \mathbf{e})[\mathbf{p}_y - \mathbf{p}_0] + \tilde{\boldsymbol{\varepsilon}}_y - \tilde{\boldsymbol{\varepsilon}}_0,$$

and the utility from purchasing Generic brand is 0. Hence, I can define $\mathbf{z}_y = \mathbf{p}_y - \mathbf{p}_0$ and $\boldsymbol{\varepsilon}_y = \tilde{\boldsymbol{\varepsilon}}_y - \tilde{\boldsymbol{\varepsilon}}_0$, $y = 1, 2, 3, 4$, where $\mathbf{p}_0$ is the price of Generic margarine.

Given that I am considering margarine products, it is not surprising that the support for $\mathbf{z}_y$ is far from being full. In particular, $\max_y \max_i z_y^{(i)} = 0.78$ and $\min_y \min_i z_y^{(i)} = -0.15$. At the same time, there is still variation in relative prices $\mathbf{z}_y$ and income $\mathbf{d}$. This variation allows me to recover β without specifying the distribution of $\boldsymbol{\varepsilon}$.

In the current application, I use a minimal amount of information: there are only two covariates. If one has more demographic and product data, it can be easily incorporated into the current framework via $\mathbf{w}$. For instance, $\mathbf{w}$ may contain nonprice marketing variables, packet size dummies, saturated fat content, household size, age of the household head, household location (e.g. zip-code).

Parametric Estimation

First, I assume the most common parametric specification for the random intercept – multinomial logit. Formally, I estimate the following specification for normalized utility:

$$\mathbb{1}\,(\, y \neq 0\,)\,[\gamma_y + (\beta_0 + \beta_1 \mathbf{d} + \mathbf{e})\mathbf{z}_y] + \alpha \boldsymbol{\varepsilon}_y,$$

where $\{\boldsymbol{\varepsilon}_y\}_{y=0}^4$ are i.i.d. Gumbel across y that are also independent from $\mathbf{x} = (\mathbf{d}, \mathbf{z})$; $\mathbf{e}$ is a standard normal random variable. (Parameter α captures the scale of $\boldsymbol{\varepsilon}_y$ since the variance of $\mathbf{e}$ is set to 1.) Although, price $\mathbf{p}_y$ is probably correlated with unobserved part of the utility $\tilde{\boldsymbol{\varepsilon}}_y$ (e.g., unobserved quality), the price difference $\mathbf{z}_y = \mathbf{p}_y - \mathbf{p}_0$ may be independent from $\tilde{\boldsymbol{\varepsilon}}_y - \tilde{\boldsymbol{\varepsilon}}_0$.

The estimates of β_0 and β_1 are $\bar{\beta}_0 = -6331.94$ (standard error= 17.19) and $\bar{\beta}_1 = -19.69$ (standard error= 514.48), respectively. As expected, the sign of $\bar{\beta}_0$ is negative. The coefficient in front of the income variable, $\bar{\beta}_1$, is negative and not significant at the 5 percent significance level. Although income does not matter much, the overall sensitivity to prices (mostly captured by $\bar{\beta}_0$ in this case) is substantial. The effect of income on marginal disutility from the price increase is not surprising given that margarine constitutes a small share of household expenditures on groceries.[19]

Semiparametric Estimation

Next, I apply the estimator proposed in Section 4. Formally, I estimate the following specification for normalized utility:

$$\mathbb{1}\,(\,y \neq 0\,)\,[(\beta_0 + \beta_1 \mathbf{d} + \mathbf{e})\mathbf{z}_y + \boldsymbol{\varepsilon}_y],$$

where $\mathbf{e}$ is a standard normal random variable. The random intercept $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_y)_{y=1,2,3,4}$ is assumed to be independent from $\mathbf{x}$. There are *no* other restrictions on the joint distribution of $\boldsymbol{\varepsilon}$. This specification nests the logit specification estimated in the previous section. Hence, if the assumptions of multinomial logit are correct, then the results of parametric and semiparametric estimators should not differ much.

The estimates of β_0 and β_1 are $\hat{\beta}_0 = -39.1$ (standard error$= 43.8$) and $\hat{\beta}_1 = -16.7 \times 10^{-3}$ (standard error$= 3.97 \times 10^{-6}$). [20] Similar to the multinomial logit estimator, the sign of $\hat{\beta}_0$ is negative. The coefficient in front of the income variable is negative and significant at the 5 percent significance level. However, the maximal value that $\hat{\beta}_1 \mathbf{d}$ can take in the sample is substantially smaller than $\hat{\beta}_0$ ($\max_i \left(\mathbf{d}^{(i)}\hat{\beta}_1/\hat{\beta}_0\right) = 0.055$, standard error$= 0.062$). The latter indicates that, similarly to the fully parametric specification, income does not affect marginal disutility from price increase much. However, the estimate of β_0 is substantially lower than the one in the fully parametric case. This indicates that consumers may be less sensitive to price changes than one would think after estimating the logit-type model.

Interestingly, the difference between the estimates obtained using the fully parametric

[19] E.g., in UK households spend about one percent of their grocery expenditures on margarine and butter (Griffith et al., 2018).

[20] I use the tensor product of the 4-th degree Chebyshev polynomials for d and the 1-st degree Chebyshev polynomials for every z_y.

logit estimator $\bar{\beta}$ and my semiparametric estimator $\hat{\beta}$ is substantial (e.g., $\bar{\beta}_1/\hat{\beta}_1 > 10^3$). That is, the parametric estimator overestimates the magnitude of the agents sensitivity to relative price changes of margarine. This suggests that the multinomial logit structure most likely fails to hold, emphasizing the importance of semiparametric estimation.[21]

7. Conclusion

This paper shows that commonly used exclusion restrictions and richness assumptions about the distribution of some unobservables may lead to full nonparametric identification in discrete outcome models even when covariates are bounded. The proposed identification framework extends the results from a large literature that uses special covariates with full support to environments where such full-support covariates are not available. It also leads to an asymptotically normal estimator of the finite-dimensional parameters of the model.

References

Allen, R. and Rehbeck, J. (2020). Identification of random coefficient latent utility models. *arXiv preprint arXiv:2003.00276*.

Allenby, G. M. and Rossi, P. E. (1991). Quality perceptions and asymmetric switching between brands. *Marketing science*, 10(3):185–204.

Andrews, D. W. K. (2011). Examples of L2-complete and boundedly-complete distributions. Discussion Paper 1801, Cowles Foundation.

Bajari, P., Hong, H., and Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. *Econometrica*, 78(5):1529–1568.

Benoit, D. F., Van Aelst, S., and Van den Poel, D. (2016). Outlier-robust bayesian multinomial choice modeling. *Journal of Applied Econometrics*, 31(7):1445–1466.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

[21] This empirical finding is in line with the simulation results, presented in Appendix B.

Berry, S., Levinsohn, J., and Pakes, A. (2004). Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of political Economy*, 112(1):68–105.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262.

Berry, S. T. and Haile, P. A. (2009). Nonparametric identification of multinomial choice demand models with heterogeneous consumers. Technical report, National Bureau of Economic Research.

Berry, S. T. and Haile, P. A. (2014). Identification in differentiated products markets using market level data. *Econometrica*, 82(5):1749–1797.

Berry, S. T. and Haile, P. A. (2020). Nonparametric identification of differentiated products demand using micro data. Technical report, National Bureau of Economic Research.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669.

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679.

Brown, L. D. (1986). *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Lecture notes – Monograph series*. Institute of Mathematical Statistics.

Canay, I. A., Santos, A., and Shaikh, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559.

Chen, S., Khan, S., and Tang, X. (2016). Informational content of special regressors in heteroskedastic binary response models. *Journal of econometrics*, 193(1):162–182.

Chen, X. and Liao, Z. (2014). Sieve m inference on irregular parameters. *Journal of Econometrics*, 182(1):70–86.

Chernozhukov, V. and Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261.

Chernozhukov, V., Imbens, G. W., and Newey, W. K. (2007). Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4–14.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

d'Haultfoeuille, X. (2011). On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471.

Dunker, F., Hoderlein, S., Kaido, H., et al. (2017). Nonparametric identification of random coefficients in endogenous and heterogeneous aggregate demand models. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Fox, J. T. (2020). A note on nonparametric identification of distributions of random coefficients in multinomial choice models. Technical report.

Fox, J. T. and Gandhi, A. (2016). Nonparametric identification and estimation of random coefficients in multinomial choice models. *The RAND Journal of Economics*, 47(1):118–139.

Fox, J. T., il Kim, K., Ryan, S. P., and Bajari, P. (2012). The random coefficients logit model is identified. *Journal of Econometrics*, 166(2):204–212.

Fox, J. T. and Lazzati, N. (2017). A note on identification of discrete choice models for bundles and binary games. *Quantitative Economics*, 8(3):1021–1036.

Fox, J. T., Yang, C., and Hsu, D. H. (2018). Unobserved heterogeneity in matching games. *Journal of Political Economy*, 126(4):1339–1373.

Gautier, E. and Hoderlein, S. (2015). A triangular treatment effect model with random coefficients in the selection equation. *arXiv preprint arXiv:1109.0362*.

Gautier, E. and Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81(2):581–607.

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review*, 97(3):713–744.

Griffith, R., Nesheim, L., and O'Connell, M. (2018). Income effects and the welfare consequences of tax in differentiated product oligopoly. *Quantitative Economics*, 9(1):305–341.

Heckman, J. (1990). Varieties of selection bias. *The American Economic Review*, 80(2):313–318.

Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216.

Ichimura, H. and Thompson, T. S. (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86(2):269–295.

Kashaev, N. and Salcedo, B. (2021). Discerning solution concepts for discrete games. *Journal of Business & Economic Statistics*, 39(4):1001–1014.

Kleiber, C. and Stoyanov, J. (2013). Multivariate distributions and the moment problem. *Journal of Multivariate Analysis*, 113:7–18.

Kline, B. (2016). The empirical content of games with bounded regressors. *Quantitative Economics*, 7(1):37–81.

Lewbel, A. (1998). Semiparametric latent variable model estimation with endogenous or mismeasured regressors. *Econometrica*, pages 105–121.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97(1):145–177.

Lewbel, A., Yan, J., Zhou, Y., et al. (2021). Semiparametric identification and estimation of multinomial discrete choice models using error symmetry. Technical report, Boston College Department of Economics.

Magnac, T. and Maurin, E. (2007). Identification and information in monotone binary models. *Journal* *of Econometrics*, 139(1):76–104.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, 27(3):313–333.

Manski, C. F. (1988). Identification of binary response models. *Journal of the American statistical Association*, 83(403):729–738.

Mattner, L. et al. (1993). Some incomplete but boundedly complete location families. *The Annals of Statistics*, 21(4):2158–2162.

Matzkin, R. L. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica: Journal of the Econometric Society*, pages 239–270.

Matzkin, R. L. (2007). Heterogeneous choice. *Econometric Society Monographs*, 43:75.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy*, 9(4):513–548.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430.

Shen, X. and Shi, J. (2005). Sieve likelihood ratio inference on general parameter space. *Science in China Series A: Mathematics*, 48(1):67–78.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1):147–165.

Thompson, T. S. (1989). Identification of semiparametric discrete choice models.

A. Proofs

I first establish identification of a more general model but without covariates w. This result will be used to prove the propositions from the main text. Assume that each instance of the environment is characterized by an endogenous outcome $\mathbf{y}$ from a known finite set Y, a vector of observed exogenous characteristics $\mathbf{x} \in X \subseteq \mathbb{R}^{d_x}$, $d_x < \infty$, that can be partitioned into $x = (d, z)$, and a vector of unobserved indexes $\mathbf{s} \in S \subseteq \mathbb{R}^{d_s}$.

Assumption 6 (Data) There exists $Y^* \subseteq Y$ such that the analyst observes (can consistently estimate) $\mu(y|x) = \Pr(\mathbf{y} = y|\mathbf{x} = x)$ for all $x \in x$ and $y \in Y^*$.

Assumption 7 There exists $h_0 : Y^* \times S \to [0,1]$, such that $\Pr(\mathbf{y} = y|\mathbf{x} = x, \mathbf{s} = s) = h_0(y, s)$, for all $y \in Y^*$, $x \in X$, and $s \in S$.

Assumption 7 is an exclusion restriction that requires $\mathbf{d}$ and $\mathbf{z}$ to affect distribution over outcomes in Y^* only via the distribution of $\mathbf{s}$.

Assumption 8 (Bounded completeness) There exists $X' \subseteq X$ such that the family of distributions $\left\{F_{\mathbf{s}|\mathbf{x}}(\cdot|x), x \in X'\right\}$ is boundedly complete.

Proposition A.1 *Under Assumptions 6-8, h_0 is identified from μ up to $F_{\mathbf{s}|\mathbf{x}}$.*

Proof. Fix some $y \in Y^*$. Under Assumption 7, I have the following integral equation

$$\forall x \in X \ : \ \mu(y|x) = \int_S h(y^*, s) dF_{\mathbf{s}|\mathbf{x}}(s|x).$$

Suppose that there exists h with $h(y^*, s) \neq h_0(y^*, s)$ for all s in some nonzero-measure set S' such that

$$\forall x \in X \ : \ \mu(y|x) = \int_S h(y^*, s) dF_{\mathbf{s}|\mathbf{x}}(s|x) = \int_S h_0(y^*, s) dF_{\mathbf{s}|\mathbf{x}}(s|x).$$

This implies that the nonzero function $h(y, \cdot) - h_0(y, \cdot)$ integrates to 0 for all $x \in X'$. The latter contradicts to Assumption 8. The fact that the choice of $y \in Y^*$ was arbitrary completes the proof. ■

A.1. Nonparametric Identification

Given a collection of random variables $\{\boldsymbol{\xi}\}_{i=1,\ldots,d}$, $d<\infty$, I say that $\boldsymbol{\xi}_i$ is redundant if there exists $j\neq i$ such that $\boldsymbol{\xi}_i=\boldsymbol{\xi}_j$ a.s.. Nonredundant elements of $\{\boldsymbol{\xi}\}_{i=1,\ldots,d}$ is the largest subset of $\{\boldsymbol{\xi}\}_{i=1,\ldots,d}$ such that non of its elements are redundant.

Assumption 9 (i) The latent $\mathbf{s}=(\mathbf{s}_i)_{i=1,\ldots,d_s}$ satisfies

$$\mathbf{s}_i=\mathbf{z}_i[\beta_{0,i}+\beta_{1,i}\mathbf{d}_i+\mathbf{e}_i] \text{ a.s.}$$

where $\beta_{0,i}$ and $\beta_{1,i}$ are some unknown parameters such that $\beta_{1,i}\neq 0$ for all $i=1,\ldots,d_s$;

(ii) Nonredundant elements of $\{\mathbf{e}_i\}_{i=1,\ldots,d_s}$ are mean-zero and variance-one independent random variables that are independent of $\mathbf{x}$;

(iii) $h_0(y^*,\cdot)$ has bounded derivatives up to order κ and $\partial^l_{s_i^l}h_0(y^*,\cdot)|_{s=0}\neq 0$ for all $l\leq\kappa$ and all $i=1,\ldots,d_s$;

(iv) The support of $\mathbf{x}$, which consists of nonredundant elements of $\{\mathbf{d}_i\}_{i=1,\ldots,d_s}$ and all of $\{\mathbf{z}_i\}_{i=1,\ldots,d_s}$, contains x^* with an open neighborhood such that $z_i^*=0$ for all $i=1,\ldots,d_s$;

(v) The sign of either $\beta_{0,i}$ or $\beta_{1,i}$ is known for every $i=1,\ldots,d_s$.

Let $\beta_0=\{\beta_{0,i}\}_{i=1}^{d_s}$ and $\beta_1=\{\beta_{1,i}\}_{i=1}^{d_s}$.

Proposition A.2 *If Assumptions 6, 7, and 9 hold, then β_0, β_1, and $\mathbb{E}\left[\mathbf{e}_i^l\right]$, $i=1,\ldots,d_s$, $0\leq l\leq\kappa$, are identified.*

Proof. Given a family $x=(x_k)_{k\in K}$ and a particular index value $k\in K$, let x_{-k} denote $(x_j)_{j\in K\setminus\{k\}}$. Fix some $i\in\{1,2,\ldots,d_s\}$ and set z_{-i} to 0. Take any $y^*\in Y^*$ from Assumption 7. To simplify notation, let $F_0:\mathbb{R}\to\mathbb{R}$ and $\eta:\mathbb{R}^2\to\mathbb{R}$ such that $F_0(t)=h_0(y^*,(0,\ldots,t,\ldots,0))$, where the only nonzero component in the second argument of h_0 is the i-th component, and $\eta(d_i,z_i)=\mu(y^*|x)$. Note that Assumption 9(iii) together with the dominated convergence theorem imply that F_0 is has bounded derivatives up to order κ.

Assumptions 7 implies that

$$\eta(d_i, z_i) = \int F_0((\beta_{0,i} + \beta_{1,i} d_i + e_i) z_i) dF_{\mathbf{e}_i|\mathbf{x}}(e|x).$$

Next, since $\mathbf{e}_i$ and $\mathbf{x}$ are independent and $h_0(y^*, \cdot)$ is κ-times differentiable with bounded derivatives, the dominated convergence theorem implies that (I dropped the subscript i from the notation)

$$\partial^l_{d^l} \eta(d, z) = \beta_1^l z^l \int \partial^l_{t^l} F_0((\beta_0 + \beta_1 d + e) z) dF_{\mathbf{e}}(e)$$

for any $l \leq \kappa$. Hence, since derivatives of $h_0(y^*, \cdot)$ are bounded, applying the dominated convergence theorem again I get that

$$\lim_{z \to 0} \frac{\partial^l_{d^l} \eta(d, z)}{z^l} = \beta_1^l \int \partial^l_{t^l} F_0(0) dF_{\mathbf{e}}(e) = \beta_1^l \partial^l_{t^l} F_0(0),$$

and, thus, $\beta_1^l \partial^l_{t^l} F_0(0)$ is identified for any $l \leq \kappa$. Similarly note that, since $h_0(y^*, \cdot)$ has bounded derivatives,

$$\partial^l_{z^l} \eta(d, 0) = \int \partial^l_{t^l} F_0(0) (\beta_0 + \beta_1 d + e)^l dF_{\mathbf{e}|\mathbf{x}}(e|x) \tag{1}$$

for every $l \leq \kappa$. Hence, since $\mathbb{E}[\,\mathbf{e}\,] = 0$ and $\beta_1 \partial_t F_0(0)$ is identified, $\beta_0 \partial_t F_0(0) = \partial_z \eta(d, 0) - \beta_1 \partial_t F_0(0) d$ is also identified. Thus, we can identify β_0/β_1 and learn the sign of β_1 from Assumption 9(v). For $l = 2$, since $\mathbb{E}[\,\mathbf{e}\,] = 0$ and $\mathbb{E}[\,\mathbf{e}^2\,] = 1$, we also can derive that

$$\partial^2_{z^2} \eta(d, 0) = \int \partial^2_{t^2} F_0(0) (\beta_0 + \beta_1 d + e)^2 dF_{\mathbf{e}|\mathbf{x}}(e|x) = \partial^2_{t^2} F_0(0) \left[(\beta_0 + \beta_1 d)^2 + 1 \right].$$

Hence, $\partial^2_{z^2} \eta(d, 0) = \beta_1^2 \partial^2_{t^2} F_0(0) \left[(\beta_0/\beta_1 + d)^2 + 1/\beta_1^2 \right].$ As a result, since we identified β_0/β_1 and $\beta_1^2 \partial^2_{t^2} F_0(0)$ in the previous steps,

$$1/\beta_1^2 = \frac{\partial^2_{z^2} \eta(d, 0)}{\beta_1^2 \partial^2_{t^2} F_0(0)} - (\beta_0/\beta_1 + d)^2$$

is identified. Since I already identified the sign of β_1 and β_0/β_1, I can identify β_0 and β_1. Moreover, I identify $\partial^l_{t^l} F_0(0)$ for all $l \leq \kappa$.

To identify all moments of $\mathbf{e}$ up to order κ, I use Equation (1) to derive the following recursive equation

$$\mathbb{E}\left[\mathbf{e}^l\right] = \frac{\partial^l_{z^l}\eta(d,0)}{\partial^l_{t^l}F_0(0)} - \sum_{k=1}^{l}\binom{l}{k}(\beta_0+d)^k\mathbb{E}\left[\mathbf{e}^{l-k}\right].$$

Going back to the original notation, I identify $\beta_{0,i}$, $\beta_{1,i}$, and $\mathbb{E}\left[\mathbf{e}_i^l\right]$, $0 \leq l \leq \kappa$. The conclusion of the proposition then follows from the fact that the choice of i was arbitrary. $\blacksquare$

Note that Proposition A.2 allows $\{\mathbf{z}_i\}_{i=1,\ldots,d_v}$ and nonredundant elements of $\{\mathbf{e}_i\}_{i=1,\ldots,d_v}$ and $\{\mathbf{d}_i\}_{i=1,\ldots,d_v}$ to have different cardinality. If the cardinality of nonredundant elements of $\{\mathbf{e}_i\}_{i=1,\ldots,d_v}$ and $\{\mathbf{d}_i\}_{i=1,\ldots,d_v}$ is the same, then the assumption that $\{\mathbf{e}_i\}_{i=1,\ldots,d_v}$ are independent can be relaxed. In this case, using a similar strategy, one can identify recursively $\mathbb{E}\left[\prod_{i\in I}\mathbf{e}_i^{\kappa_i}\right]$ for all possible $I \subseteq \{1,\ldots,d_v\}$ and set of nonnegative integers $\{\kappa_i\}_{i\in I}$ such that $\sum_{i\in I}\kappa_i \leq \kappa$. For instance, if $d_v = 2$, then for $F(v) = h(y^*,(v_1,v_2))$ I have that

$$\eta(d,z) = \int_{\mathbb{R}^2} F((\beta_{0,1}+\beta_{1,1}d_1+e_1)z_1,(\beta_{0,2}+\beta_{1,2}d_2+e_2)z_2)dF_{\mathbf{e}}(e).$$

Thus, given that β_0 and β_1 are already identified, we can identify, $\partial^2_{t_1,t_2}F(0)$ since

$$\lim_{\|z\|\to 0}\frac{\partial^2_{d_1,d_2}\eta(d,z)}{z_1z_2} = \beta_{1,1}\beta_{1,2}\int\partial^2_{t_1,t_2}F(0)dF_{\mathbf{e}}(e) = \beta_{1,1}\beta_{1,2}\partial^2_{t_1,t_2}F(0).$$

As a result, the partial derivative with respect to z_1 and z_2

$$\partial^2_{z_1,z_2}\eta(d,0) = \int_{\mathbb{R}^2}(\beta_{0,1}+\beta_{1,1}d_1+e_1)(\beta_{0,2}+\beta_{1,2}d_2+e_2)\partial^2_{t_1,t_2}F(0)dF_{\mathbf{e}}(e)$$

identifies $\mathbb{E}\left[\mathbf{e}_1\mathbf{e}_2\right]$. Similarly, one can identify $\mathbb{E}\left[\mathbf{e}_1^{\kappa_1}\mathbf{e}_2^{\kappa_2}\right]$ for all possible positive integers $\{\kappa_i\}_{i=1,2}$ such that $\sum_{i=1,2}\kappa_i \leq \kappa$.

Normal Random Coefficient

Assumption 10 (i) The latent $\mathbf{s} = (\mathbf{s}_i)_{i=1,\ldots,d_s}$ satisfies

$$\mathbf{s}_i = \mathbf{z}_i[\beta_{0,i}+\beta_{1,i}\mathbf{d}_i+\mathbf{e}_i] \text{ a.s.},$$

where $\beta_{0,i}$ and $\beta_{1,i}$ are some unknown parameters such that $\beta_{1,i} \neq 0$ for all $i = 1, \ldots, d_s$;

(ii) $\{\mathbf{e}_i\}_{i=1,\ldots,d_s}$ are i.i.d. standard normal random variables that are independent of $\mathbf{x}$;

(iii) The support of $(\mathbf{d}, \mathbf{z})$ contains an open ball;

(iv) The sign of either $\beta_{0,i}$ or $\beta_{1,i}$ is known for every $i = 1, \ldots, d_s$.

The only support restriction is imposed on $\mathbf{d}$ and $\mathbf{z}$ (Assumption 10(iii)). Assumptions 10(i)-(iii) are sufficient for Assumption 8 since the family of normal distributions indexed by the mean is complete as long as the parameter space for the mean contains an open ball.

Let $d_{-i} = (d_k)_{k \neq i}$. For a fixed $y^* \in Y^*$, d_{-i} and z, let $\eta : D_{i|d_{-i},z} \to [0,1]$ be such that for $x = ((d_i, d_{-i}), z)$, $\eta(d_i) = \mu(y^*|x)$.

Assumption 11 For every $i = 1, 2, \ldots, d_s$, there exists $y^* \in Y^*$ and $z_i \in Z_i \setminus \{0\}$ such that $\eta(\cdot)$ is neither an exponential nor an affine function.

Proposition A.3 *Suppose that Assumptions 6, 7, 10, and 11 hold. Then h_0, β_0, and β_1 are identified.*

Proof. Note that h_0 is identified up to β_0 and β_1 because of completeness of the family of normal distributions and Proposition A.1. Hence, I only need to show that β_0 and β_1 are identified. Fix some $i \in \{1, 2, \ldots, d_s\}$, z_{-i}, and d_{-i} in the support. Take y^* from Assumption 11. To simplify notation, let $F_0 : \mathbb{R} \to \mathbb{R}$ and $\eta : \mathbb{R}^2 \to \mathbb{R}$ be functions such that

$$F_0(s_i) = \int_{\mathbb{R}^{d_s - 1}} h_0(y^*, s) \prod_{k \neq i} \frac{\phi\left(s_k/z_k - \beta_{0,k} - \beta_{1,k} d_k\right)}{z_k} ds_k,$$

where $\phi(\cdot)$ is the standard normal p.d.f., and $\eta(d_i, z_i) = \mu(y^*|d, z)$.

Assumptions 7 and 10 imply that $\eta(d_i, z_i) = \int_{\mathbb{R}} F_0(s_i)\phi(s_i/z_i - \beta_{0,i} - \beta_{1,i} d_i) ds_i/z_i$. After some rearrangements and dropping subscript i from the notation, I get

$$\tilde{\eta}(d, z) = \int_{\mathbb{R}} F_0(s)\phi(s/z - \beta_0 - \beta_1 d) ds, \tag{2}$$

where $\tilde{\eta}(d, z) = z\eta(d, z)$.

Next, note that since $\partial^2_{x^2}\phi(x) = -\phi(x) - x\partial_x\phi(x)$ the following system of equations holds[22]

$$\begin{aligned}
\partial_d\tilde{\eta}(d,z) &= -\beta_1 \int F_0(t)\partial_x\phi(t/z - \beta_0 - \beta_1 d)dt,\\
\partial^2_{d^2}\tilde{\eta}(d,z) &= \beta_1^2 \int F_0(t)\partial^2_{x^2}\phi(t/z - \beta_0 - \beta_1 d)dt\\
&= -\beta_1^2\tilde{\eta}(d,z) - \beta_1(\beta_0 + \beta_1 d)\partial_d\tilde{\eta}(d,z) - \beta_1^2 \int tF_0(t)\partial_x\phi(t/z - \beta_0 - \beta_1 d)dt/z.
\end{aligned}$$

Moreover, $\partial_z\tilde{\eta}(d,z) = -\int F_0(t)t\partial_x\phi(t/z - \beta_0 - \beta_1 d)dt/z^2$. Hence,

$$\partial^2_{d^2}\tilde{\eta}(d,z) = -\beta_1^2\tilde{\eta}(d,z) - \beta_1(\beta_0 + \beta_1 d)\partial_d\tilde{\eta}(d,z) + \beta_1^2 z\partial_z\tilde{\eta}(d,z).$$

Equivalently,

$$\frac{\beta_0}{\beta_1} = \frac{z\partial_z\tilde{\eta}(d,z) - \tilde{\eta}(d,z)}{\partial_d\tilde{\eta}(d,z)} - d - \frac{\partial^2_{d^2}\tilde{\eta}(d,z)}{\partial_d\tilde{\eta}(d,z)}\frac{1}{\beta_1^2}.$$

Replacing $\tilde{\eta}(d,z)$ by $z\eta(d,z)$, I get

$$\frac{\beta_0}{\beta_1} = \frac{z\partial_z\eta(d,z) - d\partial_d\eta(d,z)}{\partial_d\eta(d,z)} - \frac{\partial^2_{d^2}\eta(d,z)}{\partial_d\eta(d,z)}\frac{1}{\beta_1^2}. \tag{3}$$

Thus, β_0/β_1 is identified up to β_1^2. Differentiating the last equation with respect to d leads to the following equation

$$\frac{1}{\beta_1^2} = \partial_d\left[\frac{z\partial_z\eta(d,z) - d\partial_d\eta(d,z)}{\partial_d\eta(d,z)}\right]/\partial_d\left[\frac{\partial^2_{d^2}\eta(d,z)}{\partial_d\eta(d,z)}\right]. \tag{4}$$

Hence, if

$$\partial_d\left[\frac{\partial^2_{d^2}\eta(d,z)}{\partial_d\eta(d,z)}\right] \neq 0 \tag{5}$$

for *some* d and z, then β_1^2 is identified. Suppose this is not the case. That is, for all d and z $\partial_d\left[\frac{\partial^2_{d^2}\eta(d,z)}{\partial_d\eta(d,z)}\right] = 0$. Equivalently, $\partial^2_{z^2}\left[\log(\partial_d\eta(d,z))\right] = 0$ for all d and z. The latter would imply that either $\eta(d,z) = K_1(z)e^{K_3(z)d} + K_2(z)$ or $\eta(d,z) = K_1(z)d + K_2(z)$ for some

[22] I can differentiate under the integral sign since (i) h_0 being bounded implies that F_0 is bounded, (ii) all derivatives of the standard normal p.d.f. are bounded.

functions $K_k(\cdot)$, $k = 1, 2, 3$. Since it is assumed that $\eta(\cdot, z)$ is neither an exponential nor an affine function on some open set, I can conclude that for some d and z Equation (5) is satisfied. Thus, β_1^2 is identified (hence, $|\beta_1|$ is also identified). Hence, I identify β_0/β_1. If $\beta_0/\beta_1 = 0$, then the sign of β_1 is identified from Assumption 10(iv). If $\beta_0/\beta_1 \neq 0$, then the sign of either β_1 or β_0 is identified from Assumption 10(iv). Knowing the sign of, say, β_0 and β_0/β_1 identifies β_1 and β_0. Going back to the original notation I identify $\beta_{1,i}$ and $\beta_{0,i}$. The conclusion of the proposition then follows from the fact that the choice of i was arbitrary.

Note that for identification of β_1 and β_0, I do not need to exclude all exponential functions of d, since instead of differentiating Equation (3) with respect to d, I can differentiate it with respect to z. For the identification result to hold it suffices to exclude functions of the form $\eta(d, z) = K_1(z)e^{K_2 d} + K_3(z)$ or $\eta(d, z) = K_1(z)d + K_3$, where $K_1(\cdot)$ and $K_2(\cdot)$ are some functions of z, and K_3 is a constant. ■

A.2. Proof of Propositions 3.1, 3.2, and 3.4

In the previous section, I stated and proved two general identification results (Propositions A.2 and A.3). Next I will apply these results to a multinomial choice model studied in the main text of the paper.

Fix some arbitrary $w \in W$. To prove Propositions 3.1 and 3.4(i), I use Propositions A.2 and A.3. Both propositions require Assumptions 6 and 7. Assumptions 6 is implied by Assumption 1 for $Y^* = \{0\}$. Assumption 7 is satisfied in Propositions 3.1 with $h(0, s) = F_{\varepsilon|\mathbf{w}}(0, \ldots, s, \ldots, 0)$, where the the only nonzero component corresponds to $\bar{y}$ from Assumption 3(ii). To show validity of Assumption 7 in Proposition A.3, note that under Assumption 3.(iii) or Assumption 4.(ii) there exists z^* and $\{\lambda_y\}_{y=1}^J$ with some open neighbourhood such that $z^*_{y'} = \lambda_{y'} z^*_1$ for all $y' \in Y$ with $\min_{y'} \lambda_{y'} > 0$. Note that since $\mathbf{e}$ and $\mathbf{z}$ are independent conditional on $\mathbf{w}$, I have that for $x^* = (d^*, z^*, w)$

$$\mu(0|x^*) = \textstyle\int_{\mathbb{R}} F_{\varepsilon|\mathbf{w}}(-z_1^*(\beta_0(w) + \beta_1(w)d^* + e), \ldots, -\lambda_J z_1^*(\beta_0(w) + \beta_1(w)d^* + e)|w) dF_{\mathbf{e}|\mathbf{w}}(e|w). \quad (6)$$

Hence, Assumptions 7 is satisfied for $h(0, s) = F_{\varepsilon|\mathbf{w}}(s, \lambda_2 s, \cdot, \lambda_J s|w)$. The rest of assumptions follow from Assumption 3 or Assumption 4, except for identification of the sign of β_0 or

β_1. However, I can identify the sign of $\beta_1(w)$ from Equation (6) since $F_{\varepsilon|\mathbf{w}}(\cdot|w)$ is weakly monotone. As a result, I can identify $\beta_0(w)$, $\beta_1(w)$, and $\mathbb{E}\left[\mathbf{e}^l|\mathbf{w}=w\right]$, $0 \le l \le \kappa$ (if $\mathbf{e}$ is standard normal then we already know its distribution). The fact that the choice of w was arbitrary completes the proof.

To prove Propositions 3.2 and 3.4(ii), note that since β_0, β_1, and $F_{\mathbf{e}|\mathbf{x}}$ are identified (either from its moments or because it is standard normal), I know the distribution of $\mathbf{v} = \beta_0(\mathbf{w}) + \beta_1(\mathbf{w})\mathbf{d} + \mathbf{e}$. Moreover, $F_{\mathbf{v}|\mathbf{x}}$ constitutes a boundedly complete family either by the assumption in Proposition 3.2 or by normality of $\mathbf{e}$ and continuity of $\mathbf{d}$ in an open ball (Brown, 1986). Hence, since

$$\Pr(\mathbf{y}=0|\mathbf{x}=x) = \int_{\mathbb{R}} F_{\varepsilon|\mathbf{w}}(-z_1 v, \dots, -z_J v|w) dF_{\mathbf{e}|\mathbf{w}}(v - \beta_0(w) - \beta_1(w)d|w) =$$
$$= \int_{\mathbb{R}} \tilde{g}(z,w,v) dF_{\mathbf{e}|\mathbf{w}}(v - \beta_0(w) - \beta_1(w)d|w)$$

and Assumptions 7 is satisfied, I can identify $\tilde{g}(z,w,v) = F_{\varepsilon|\mathbf{w}}(-z_2 v, \dots, -z_J v|w)$ for all z, w, v by Proposition A.1. Note that since v can take any value in

$$V_w = \{v \,:\, v = e + \beta_1(w)d + \beta_0(w), e \in E_w, d \in D_w\}$$

for any direction $-z/\|z\|$ in the support of $\mathbf{z}$ conditional on $\mathbf{w}=w$, I can recover $F_{\varepsilon|\mathbf{w}}(g|w)$ for any g such that $g = -zv/\|z\|$ for some v. That is, I identify $F_{\varepsilon|\mathbf{w}}(\cdot|w)$ over the set R_w.

A.3. Proof of Proposition 4.1

To simplify the notation, I will focus on the binary choice case.

Step 1. In this step I make several observations about p_0 and its derivatives. By definition $0 \le h_0(v) \le 1$ for all v and

$$p_0(x) = \int_{\mathbb{R}} h_0((\beta_0 + \beta_1 d + e)z_1)\phi(e)de = \int_{\mathbb{R}} h_0(v)\phi(v/z_1 - \beta_1 d - \beta_0)dv/z_1.$$

Hence, p_0 is continuously differentiable of any order. Moreover, $p_0(x) = 0$ if and only if $h(v) = 0$ for all v. The latter means that probability of picking the outside option conditional on $\mathbf{x} = x$ and $\mathbf{e} = e$ equals to 0 for all e. Since $\boldsymbol{\varepsilon}_1$ is independent of $\mathbf{x}$ and $\mathbf{e}$, I have that $\boldsymbol{\varepsilon}_1 \geq -z_1(\beta_0 + \beta_1 d + e)$ with probability 1 for all e, which is not possible since $\mathbf{e}$ has full support. Thus, $p_0(x) > 0$ for all x. Similarly, one can show that $p_0(x) < 1$ for all x.

Next consider $p_1(x) = \partial_d p_0(x)$. Since $\partial_t \phi(t) = -t\phi(t)$,

$$|p_1(x)| = \left|\beta_1 \int_{\mathbb{R}} h_0(v)(v/z_1 - \beta_1 d - \beta_0)\phi(v/z_1 - \beta_1 d - \beta_0)dv/z_1\right| = \left|\beta_1 \int_{\mathbb{R}} h_0((\beta_0 + \beta_1 d + e)z_1)e\phi(e)de\right|.$$

Hence, since $0 \leq h_0(v) \leq 1$ for all v, I get that for some $C_1 < \infty$, $\sup_x |p_1(x)| \leq \beta_1 \int_{\mathbb{R}} |e|\,\phi(e)de \leq C_1$. Similarly, note that $p_2(x) = z_1 \partial_{z_1} p_0(x)$ and by the triangular inequality

$$|p_2(x)| \leq |p_0(x)| + \left|\int_{\mathbb{R}} h_0((\beta_0 + \beta_1 d + e)z_1)e\phi(e)(\beta_0 + \beta_1 d + e)de\right|.$$

Hence, given bounded support of x, I can conclude that $\sup_x |p_2(x)|$ is also finite. Repeating the above steps, one can show that all higher order partial derivatives of p_0 are bounded.

Step 2. Note that in the proof of Proposition 3.4 we used derivatives of $\eta(d, z_1)$ to identify βs. In particular, we can take $\eta(d^{**}, z_1^{**}) = \mu(0|x^{**})$, where $x^{**} = (d^{**}, (\lambda_y z_1^{**}))_y, w)$ and $\lambda_y = z_{2,y}^{**}/z_{2,1}^{**}$. As a result, $\partial_{z_1}\eta(d^{**}, z_1^{**}) = \sum_y \lambda_y \partial_{z_y}\mu(0|x^{**})$. Since $\lambda_y = z_{2,y}^{**}/z_{2,1}^{**}$, I get that $z_1^{**}\partial_{z_1}\eta(d^{**}, z_1^{**}) = \sum_y z_y^{**}\partial_{z_y} \Pr(\mathbf{y} = 0|\mathbf{x} = x^{**})$. Hence, if Assumption 4(iii) is satisfied not just for one (d^{**}, z^{**}) but for all, then for all x

$$\begin{aligned}
\beta_1^2 &= \frac{\partial_{d^3}^3 p_0(x)\partial_d p_0(x) - [\partial_d p_0(x)]^2}{\sum_y z_y \partial_{d,z_y}^2 p_0(x)\partial_d p_0(x) - \sum_y z_y \partial_{z_y} p_0(x)\partial_{d^2}^2 p_0(x) - [\partial_d p_0(x)]^2},\\
\beta_0 &= \frac{\sum_y z_y \partial_{z_y} p_0(x) - d\partial_d p_0(x)}{\partial_d p_0(x)}\beta_1 - \frac{\partial_{d^2}^2 p_0(x)}{\partial_d p_0(x)}\frac{1}{\beta_1}. \qquad (7)
\end{aligned}$$

Step 3. Combining the bounds for the derivatives from Step 1, the uniform weak law of large numbers, and consistency of $\hat{p}_0$, I can deduce that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \hat{p}_{111}\left(\mathbf{x}^{(i)}\right)\hat{p}_1\left(\mathbf{x}^{(i)}\right) - \hat{p}_{11}\left(\mathbf{x}^{(i)}\right)^2 \to_p \mathbb{E}\left[p_{111}(\mathbf{x})p_1(\mathbf{x}) - p_{11}(\mathbf{x})^2\right],\\
&\frac{1}{n}\sum_{i=1}^n \hat{p}_{12}\left(\mathbf{x}^{(i)}\right)\hat{p}_1\left(\mathbf{x}^{(i)}\right) - \hat{p}_2\left(\mathbf{x}^{(i)}\right)\hat{p}_{11}\left(\mathbf{x}^{(i)}\right) - \hat{p}_1\left(\mathbf{x}^{(i)}\right)^2 \to_p \mathbb{E}\left[p_{12}(\mathbf{x})p_1(\mathbf{x}) - p_2(\mathbf{x})p_{11}(\mathbf{x}) - p_1(\mathbf{x})^2\right],\\
&\frac{1}{n}\sum_{i=1}^n \hat{p}_2\left(\mathbf{x}^{(i)}\right) - \mathbf{d}^{(i)}\hat{p}_1\left(\mathbf{x}^{(i)}\right) \to_p \mathbb{E}\left[p_2(\mathbf{x}) - \mathbf{d}p_1(\mathbf{x})\right],
\end{aligned}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{p}_{11}\left(\mathbf{x}^{(i)}\right) \to_p \mathbb{E}\left[\,p_{11}(\mathbf{x})\,\right], \qquad \frac{1}{n}\sum_{i=1}^{n}\hat{p}_{1}\left(\mathbf{x}^{(i)}\right) \to_p \mathbb{E}\left[\,p_{1}(\mathbf{x})\,\right].$$

Thus, Equation (7) and the continuous mapping theorem imply that $\hat{\beta} \to_p \beta$.

Step 4. Consider

$$\mathcal{G}_n = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \hat{p}_{111}\left(\mathbf{x}^{(i)}\right)\hat{p}_1\left(\mathbf{x}^{(i)}\right) - \hat{p}_{11}\left(\mathbf{x}^{(i)}\right)^2 \\ \beta_1^2\left[\hat{p}_2\left(\mathbf{x}^{(i)}\right) - \mathbf{d}^{(i)}\hat{p}_1\left(\mathbf{x}^{(i)}\right)\right] - \hat{p}_{11}\left(\mathbf{x}^{(i)}\right) \end{pmatrix}.$$

To prove asymptotic normality of $\mathcal{G}_n$, I will use Theorem 6 in Newey (1997). The data is assumed to be i.i.d., the outcome variable is finite and p_0 is bounded and bounded away from 0. Hence, Assumptions 1 and 4 from Newey (1997) are satisfied. Assumption 8 in Newey (1997) is assumed. Assumption 9 in Newey (1997) follows from Step 1. Finally, consider $a(p_0) = (a_1(p_0), a_0(p_0))$ with

$$a_1(p_0) = \mathbb{E}\left[\,p_{111}(\mathbf{x})p_1(\mathbf{x}) - p_{11}(\mathbf{x})^2\,\right], \qquad a_2(p_0) = \mathbb{E}\left[\,\beta_1^2[p_2(\mathbf{x}) - \mathbf{d}p_1(\mathbf{x})] - p_{11}(\mathbf{x})\,\right].$$

The directional derivative of a at p_0 in direction g_0 is then $D(g_0) = (D_1(g_0), D_2(g_0))$ with

$$D_1(g_0) = \mathbb{E}\left[\,p_{111}(\mathbf{x})g_1(\mathbf{x}) + g_{111}(\mathbf{x})p_1(\mathbf{x}) - 2p_{11}(\mathbf{x})g_{11}(\mathbf{x})\,\right], \quad D_2(g_0) = \mathbb{E}\left[\,\beta_1^2[g_2(\mathbf{x}) - \mathbf{d}g_1(\mathbf{x})] - g_{11}(\mathbf{x})\,\right].$$

Applying integration by parts several times and using the fact that $f_{\mathbf{x}}$ and its partial derivatives vanish at the boundary of the support of $\mathbf{x}$ (Assumption 5(iii)), I get

$$\begin{aligned}
\mathbb{E}\left[\,p_{111}(\mathbf{x})g_1(\mathbf{x})\,\right] &= -\mathbb{E}\left[\,\partial_{z_1}[p_{111}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x})]g_0(\mathbf{x})/f_{\mathbf{x}}(\mathbf{x})\,\right], \\
\mathbb{E}\left[\,p_{1}(\mathbf{x})g_{111}(\mathbf{x})\,\right] &= -\mathbb{E}\left[\,\partial^3_{z_1^3}[p_{1}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x})]g_0(\mathbf{x})/f_{\mathbf{x}}(\mathbf{x})\,\right], \\
\mathbb{E}\left[\,p_{11}(\mathbf{x})g_{11}(\mathbf{x})\,\right] &= \mathbb{E}\left[\,\partial^2_{z_1^2}[p_{11}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x})]g_0(\mathbf{x})/f_{\mathbf{x}}(\mathbf{x})\,\right], \\
\mathbb{E}\left[\,\mathbf{d}g_1(\mathbf{x})\,\right] &= -\mathbb{E}\left[\,(f_{\mathbf{x}}(\mathbf{x}) + \mathbf{d}\partial_{z_1}f_{\mathbf{x}}(\mathbf{x}))g_0(\mathbf{x})/f_{\mathbf{x}}(\mathbf{x})\,\right], \\
\mathbb{E}\left[\,g_{11}(\mathbf{x})\,\right] &= \mathbb{E}\left[\,\partial^2_{z_1^2}f_{\mathbf{x}}(\mathbf{x})g_0(\mathbf{x})/f_{\mathbf{x}}(\mathbf{x})\,\right], \\
\mathbb{E}\left[\,g_2(\mathbf{x})\,\right] &= -\mathbb{E}\left[\,(f_{\mathbf{x}}(\mathbf{x}) + \mathbf{z}_1\partial_{z_2}f_{\mathbf{x}}(\mathbf{x}))g_0(\mathbf{x})/f_{\mathbf{x}}(\mathbf{x})\,\right].
\end{aligned}$$

As a result,

$$D_1(g_0) = -\mathbb{E}\left[\{4p_{1111}(\mathbf{x})f_{\mathbf{x}}(\mathbf{x}) + 8p_{111}(\mathbf{x})\partial_{z_1} f_{\mathbf{x}}(\mathbf{x}) + 5p_{11}(\mathbf{x})\partial^2_{z_1^2} f_{\mathbf{x}}(\mathbf{x}) + p_1(\mathbf{x})\partial^3_{z_1^3} f_{\mathbf{x}}(\mathbf{x})\}g_0(\mathbf{x})/f_{\mathbf{x}}(\mathbf{x})\right],$$
$$D_2(g_0) = \mathbb{E}\left[\{\beta_1^2[\mathbf{d}\partial_d f_{\mathbf{x}}(\mathbf{x}) - \mathbf{z}_1\partial_{z_2} f_{\mathbf{x}}(\mathbf{x})] - \partial^2_{d^2} f_{\mathbf{x}}(\mathbf{x})\}g_0(\mathbf{x})/f_{\mathbf{x}}(\mathbf{x})\right].$$

Hence, $D(g_0) = \mathbb{E}\left[\bar{v}(\mathbf{x})g_0(\mathbf{x})\right]$. Moreover, $\bar{v}$ is continuously differentiable and $\mathbb{E}\left[\bar{v}(\mathbf{x})\bar{v}(\mathbf{x})^{\mathsf{T}}\right]$ is finite and nonsingular (Assumption 5(iv)). Hence, Assumption 7 in Newey (1997) is also satisfied, thus, by Theorem 6 in Newey (1997), $\sqrt{n}\,(\mathcal{G}_n - \mathcal{G}) \to_d N(0, \tilde{V})$, where

$$\mathcal{G} = \mathbb{E}\left[\begin{array}{c} p_{111}(\mathbf{x})p_1(\mathbf{x}) - p_{11}(\mathbf{x})^2 \\ \beta_1^2\left[p_2(\mathbf{x}) - \mathbf{d}p_1(\mathbf{x})\right] - p_{11}(\mathbf{x}) \end{array}\right]$$

and $\tilde{V} = \mathbb{E}\left[\bar{v}(\mathbf{x})\bar{v}(\mathbf{x})^{\mathsf{T}}p_0(\mathbf{x})(1 - p_0(\mathbf{x}))\right]$. Moreover, I can construct a consistent estimator of $\tilde{V}$ using Theorem 6 in Newey (1997). In particular, let $\hat{a}(\hat{p}_0)$ be a sample counterpart of $a(p_0)$ and

$$\hat{\gamma} = \left(\Psi^{\mathsf{T}}\Psi\right)^{-}\sum_{i=1}^{n}\psi^K\left(\mathbf{x}^{(i)}\right)\mathbb{1}\left(\mathbf{y}^{(i)} = 0\right), \qquad \hat{A} = \partial_\gamma \hat{a}(\psi^K(z)^{\mathsf{T}}\hat{\gamma}),$$
$$\hat{Q} = \Psi^{\mathsf{T}}\Psi/n, \qquad \hat{\Sigma} = \sum_{i=1}^{n}\psi^K\left(\mathbf{x}^{(i)}\right)\psi^K\left(\mathbf{x}^{(i)}\right)^{\mathsf{T}}\left[\mathbb{1}\left(\mathbf{y}^{(i)} = 0\right) - \hat{p}_0\left(\mathbf{x}^{(i)}\right)\right]^2/n.$$

Then $\hat{\tilde{V}} = \hat{A}^{\mathsf{T}}\hat{Q}^{-}\hat{\Sigma}\hat{Q}^{-}\hat{A} \to_p \tilde{V}$.

Step 5. Combining Step 2 with the continuous mapping theorem, Slutsky's theorem, and the Delta method, implies that

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to_d \begin{pmatrix} 2\beta_1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \mathbb{E}\left[p_{12}(\mathbf{x})p_1(\mathbf{x}) - p_2(\mathbf{x})p_{11}(\mathbf{x}) - p_1(\mathbf{x})^2\right] & 0 \\ 0 & \beta_1\mathbb{E}\left[p_1(\mathbf{x})\right] \end{pmatrix}^{-1} N\left(0, \tilde{V}\right).$$

Step 5. Consistency of $\hat{V} = \hat{G}\hat{\tilde{V}}\hat{G}^{\mathsf{T}}$, where

$$\hat{G} = \begin{pmatrix} 2\hat{\beta}_1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} n^{-1}\sum_{i=1}^{n}\hat{p}_{12}\left(\mathbf{x}^{(i)}\right)\hat{p}_1\left(\mathbf{x}^{(i)}\right) - \hat{p}_2\left(\mathbf{x}^{(i)}\right)\hat{p}_{11}\left(\mathbf{x}^{(i)}\right) - \hat{p}_1\left(\mathbf{x}^{(i)}\right)^2 & 0 \\ 0 & n^{-1}\hat{\beta}_1\sum_{i=1}^{n}\hat{p}_1\left(\mathbf{x}^{(i)}\right) \end{pmatrix}^{-1},$$

follows from consistency of $\hat{\beta}$, $\hat{\tilde{V}}$, Step 3, and the continuous mapping theorem.

B. Additional Simulations

Table 3 contains results for the mean absolute deviation (MAD) of my estimator of β_1. Similar to the bias, the MAD decreases with n and is of the similar magnitude across DGPs.

Table 3 – Mean Absolute Deviation

Sample Size/DGP	DGP-0	DGP-1	DGP-2	DGP-3	DGP-4	DGP-5	DGP-L
1000	1.11	1.13	1.20	1.48	1.53	1.52	1.15
5000	0.48	0.65	0.95	1.09	1.28	1.25	0.71
10000	0.38	0.42	0.67	0.90	1.13	1.21	0.52

Next, I estimated β_1 using two maximum-likelihood estimators. The first one (Probit) is based on the assumption that ε is standard normal. The second one (Logit) is assumes that ε has a logistic distribution. The Probit estimator is correctly specified under DGP-0 and is misspecified for all other DGPs. The Logit estimator is misspecified for all DGPs except DGP-L. The results for the bias and the MAD for both estimators for $n = 1000$ are presented in Tables 4 and 5.

Table 4 – Bias and Mean Absolute Deviation of the Probit estimator

Metric/DGP	DGP-0	DGP-1	DGP-2	DGP-3	DGP-4	DGP-5	DGP-L
Bias	0.05	26.0	46.25	183.18	716.74	2197.74	25.05
MAD	0.14	26.1	46.35	183.28	716.82	2197.81	25.19

Table 5 – Bias and Mean Absolute Deviation of the Logit estimator

Metric/DGP	DGP-0	DGP-1	DGP-2	DGP-3	DGP-4	DGP-5	DGP-L
Bias	0.06	0.25	0.66	2.76	7.76	16.96	0.47
MAD	0.15	0.34	0.77	2.85	7.84	17.01	0.59

Overall, the Logit estimator outperforms the Probit estimator for all DGPs except DGP-0. As expected, since for DGP-0 and DGP-L the Probit and the Logit estimators are correctly specified, respectively, the bias and the MAD are small and both estimators perform better than my estimator (see also Table 1). However, for the rest of DGPs, these estimators perform very poorly. For instance, the bias of the Logit estimator is about 11 times bigger that the bias of my estimator for DGP-5.