Identification and Estimation of Multinomial Choice Models with Latent Special Covariates

Nail Kashaev
Department of Economics, University of Western Ontario, London, Canada

ABSTRACT
Identification of multinomial choice models is often established by using special covariates that have full support. This article 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.

ARTICLE HISTORY
Received September 2020
Accepted March 2022

KEYWORDS
Bundles; Identification at infinity; Multinomial choice; Random coefficients; Special covariate

1. Introduction
This article 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 article 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. 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 nonparametrically 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, Stock, and Stoker (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, il Kim, Ryan, and Bajari 2012). This article is also related to the literature on identification of finite-dimensional parameters and evaluate utility from choosing an alternative

\[ y = \begin{cases} 1 & \text{if } z \in Y \text{ is preferred}, \\ 0 & \text{otherwise}. \end{cases} \]

normalize the utility from alternative

\[ \text{utility} = \frac{\text{utility from alternative } y}{\text{utility from alternative } z}, \]

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 nonparametrically 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 article provide a more solid econometric foundation to the models with at least one normally distributed random coefficient (e.g., Nevo 2000).

The article 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 models 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 \( f \) 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

\[ y = 0 \text{ to } 0. \]

The random utility from choosing an alternative \( y \neq 0 \) is

\[ z_y[\beta_0(w) + \beta_1(w)d + e] + \varepsilon_y, \]

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

The random coefficient \( \beta_0(w) + \beta_1(w)d + e \) represents a specific heterogeneous tastes associated with the product characteristic \( z_y \) (i.e., the marginal utility from the product characteristic \( z_y \)). This specification of random coefficients is common in applied work (see, for instance, Berry, Levinsohn, and Pakes 1995; Nevo 2000, 2001; Berry, Levinsohn, and Pakes 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 \( d \) and \( \beta_1(w) \) is scalar without loss of generality since if \( d \) is a vector, then all components of it but one can be absorbed by \( w \). In this case, one would need to use variation in those absorbed components to identify the coefficients in front of them. Similarly to the existing treatment of random coefficients model, I assume that the random coefficients in front of \( z_y \) are the same for each alternative \( y \). However, I do not impose sign restrictions on \( \beta_0(w) + \beta_1(w)d + e \).

I start by stating two assumptions that will be used throughout the article. 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 \( p_0(x) = \Pr(y = 0|x = x) \) 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, il Kim, Ryan, and Bajari 2012). If the information on the identity of the purchases is also available, then this information (a) may improve the efficiency of an estimator; (b) can help to satisfy the assumptions needed for identification (e.g., in my empirical illustration, I use one product to identify the sign of \( \beta_1 \) and I use another one to estimate it); and (c) can be used to weaken the assumption that the random slope coefficient \( \beta_0(w) + \beta_1(w)d + e \) is the same across inside goods.

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

i. \( e \) is conditionally independent of \( (e, d, z) \) conditional on \( w = w' \);

ii. \( e \) is conditionally independent of \( (d, z) \) conditional on \( w = w' \).

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

Next, I provide two nonnested sets of conditions that allow for identification of \(\beta_0\), \(\beta_1\), and the distribution of \(e\) and \(\epsilon\). In Section 3.1, I impose no parametric assumptions on latent \(e\) and \(\epsilon\) but assume some smoothness on the cdf of \(\epsilon\) and restrict the support of covariates. In Section 3.2, I identify the model when \(e\) is normally distributed, without any additional restrictions on the distribution of \(\epsilon\) and with minimal support restrictions on covariates.

3. Identification

3.1. Nonparametric Identification

Assumption 3. For all \(w \in W\)

1. Conditional on \(w = w\), \(e\) has mean zero and variance one;
2. \(\text{F}_{\epsilon|w}(\cdot|w)\) has bounded partial derivatives up to order \(\kappa\) for some \(y\) and \(\frac{\partial^{\kappa}}{\partial^\kappa} \text{F}_{\epsilon|w}(\cdot|w)|_{\epsilon=0} = 0\) for all \(l \leq \kappa\);
3. There exists \(d^*\) such that the support of \((d, z)\) conditional on \(w = w\) contains \((d^*, 0)\) with an open neighborhood.

Assumption 3(i) is a scale and location normalization. It restricts \(e\) conditional on \(w = w\) to have a finite expectation and a nonzero variance for all \(w\). Assumption 3(ii) requires the conditional distribution \(\text{F}_{\epsilon|w}\) to be sufficiently smooth in one component of \(e\) in the neighborhood of zero and have different from zero higher order partial derivatives. Since \(E[\epsilon]\) is not assumed to be zero, if, for instance, \(e\) is multivariate normal with component-wise nonzero mean, then Assumption 3(ii) is automatically satisfied. It is also generically satisfied when at least one component of \(e\) is independent of the others and has a type I extreme value distribution (Fox, il Kim, Ryan, and Bajari 2012). However, Assumption 3(ii) rules out cases when \(e\) is a constant. Another example of violation of Assumption 3(ii) is when \(\kappa\) is infinite and \(\text{F}_{\epsilon|w}\) is a polynomial function of any finite degree. (In Section 3.2, I provide an alternative result that does not restrict \(\text{F}_{\epsilon|w}\).) Assumption 3(iii) requires the support of \(z\) 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, il Kim, Ryan, and Bajari 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 \(E[\epsilon|w = w], 0 \leq l \leq \kappa\), are identified for all \(w \in W\).

Identification of \(\kappa \leq \infty\) moments of the conditional distribution of \(\epsilon\) conditional on \(w\) is often sufficient for nonparametric identification of it. For example, Assumption 7 in Fox, il Kim, Ryan, and Bajari (2012) uses the Carleman condition.\(^8\) Thus, under minimal restrictions, I can nonparametrically identify the conditional cdf \(\text{F}_{\epsilon|x}\), where \(v = \beta_0(w) + \beta_1(w)d + \epsilon\).

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

Definition 1 (Bounded completeness). The family of distributions \(\{\text{F}_{\epsilon|x}(\cdot|x), x \in X'\}\) is boundedly complete if

\[
\forall x \in X', \int_x g(t)d\text{F}_{\epsilon|x}(t|x) = 0 \implies g(v) = 0 \text{ a.s.,}
\]

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 distribution.\(^9\)

Combining bounded completeness with the identified distribution of the index \(v\), I have the following result.

Proposition 3.2. If \(\text{F}_{\epsilon|x}\) is identified and \(\{\text{F}_{\epsilon|x}(\cdot|(d, z, w)), d \in D_{(z, w)}\}\) 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 \(\{y \neq 0\} (r_\sigma + e_\sigma)\). The vector \(r = (r_\sigma)_{\sigma \in \gamma \backslash \{0\}}\) is observed covariate conditionally independent of \(e = (e_\sigma)_{\sigma \in \gamma \backslash \{0\}}\) conditional on \(w = w\) with the conditional support \(R_w = \{r \in \mathbb{R}^I : r = vz, z \in Z_w, v \in V_w\}\), where \(V_w\) is the support of \(v\) conditional on \(w = w\). In particular, \(\text{F}_{\epsilon|w}\) is identified over \(-R_w\).

The proof of Proposition 3.2 is similar to the proof of Theorem 11 in Fox, il Kim, Ryan, and Bajari (2012). The main difference is that, instead of parametric restrictions, Proposition 3.2 uses the interaction between \(d\) and \(z\).

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 \(z/\|z\|\) can cover is sufficiently rich and the support of \(e\) conditional on \(w = w\) is \(\mathbb{R}\), then \(R_w = \mathbb{R}^I\) 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.

---

\(^7\)Since, for identification and estimation, I require the average structural function \(p_0\), some forms of endogeneity (i.e., correlation between \(x\) and \(e\)) can be addressed using suitable instruments and control function residuals as in Blundell and Powell (2004) (see also Berry 1994; Berry, Levinsohn, and Pakes 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.

\(^8\)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.

\(^9\)For testability of the completeness assumptions see Canay, Santos, and Shaikh (2013).
Corollary 3.1. For all \( y \neq 0 \), let \( \epsilon_y = \theta^T w_y + \zeta_y \), where \( \theta \) and \( \zeta = (\zeta_y)_{y \in Y \setminus \{0\}} \) are random coefficients, and \( w_y \) is the vector of product-\( y \)-specific covariates. Suppose

i. The assumptions of Propositions 3.1 and 3.2 hold;
ii. \( R_w = \mathbb{R}^I \) for all \( w \in W \);
iii. \( (\theta, \zeta) \) and \( w = (w_y)_{y \in Y \setminus \{0\}} \) are independent;
iv. The support of \( w \) contains an open ball of dimensionality of \( w \);
v. \( (\theta, \zeta) \) has finite absolute moments and its distribution is uniquely determined by its moments;

then \( \beta_0, \beta_1 \), and the distribution of \( (\epsilon, \theta, \epsilon) \) are identified.

To the best of my knowledge, Corollary 3.1 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, il Kim, Ryan, and Bajari (2012), Allen and Rehbeck (2020), and Lewbel, Yan, and Zhou (2021) also allow for bounded covariates. However, they either do not fully identify the distribution of the random intercept \( \epsilon \) (Allen and Rehbeck 2020; Lewbel, Yan, and Zhou 2021) or impose parametric restrictions on it (Fox, il Kim, Ryan, and Bajari 2012).

3.2. Normal Taste Shock

Assumption 4. For all \( w \in W \)

i. Conditional on \( w = w \), \( \epsilon \) is a standard normal random variable;
ii. there exists \((d^w, z^w)\) in the interior of the support of \((d, z)\) conditional on \( w = w \) such that \( z^w_y \neq 0 \) for all \( y \in Y \);
iii. there exists \((d^{w*}, z^{w*})\) in the interior of the support of \((d, z)\) conditional on \( w = w \) such that \( p_0((d, z^{w*}, w)) \) is neither an exponential nor an affine function of \( d \) on some open set.

Assumption 4(i) requires \( \epsilon \) to be normally distributed with nonzero variance. Without nonzero variance, the assumption that \( \mathbb{E}[\epsilon]\) 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 3.3. 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 \( \beta_0 \) and \( \beta_1 \) uses the multiplicative structure of \( d \) and \( z \), and properties of the standard normal pdf. Informally, consider

\[
\beta_0(w)z + \beta_1(w)dz + ez.
\]

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 3.3(ii) follows from \( \beta_0(w) \) and \( \beta_1(w) \) being identified and \( e \) being standard normal (i.e., \( \beta_0(w) + \beta_1(w)d + e \) conditional on \( x = x \) generates a boundedly complete family of distributions).

Note that the only restriction on \( e \) needed for Proposition 3.3 is the conditional independence assumption (Assumption 2). The random intercept \( e \) 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 3.3. For all \( y \neq 0 \) let \( \epsilon_y = \theta_y(w) \), where \( \theta_y \) is a random function such that its realization \( \theta_y \) is a map from \( W \) to \( \mathbb{R} \). Suppose

i. Assumptions of Proposition 3.3 hold;
ii. \( R_w = \mathbb{R}^I \) for all \( w \in W \);
iii. \( \theta = (\theta_y)_{y \neq 0} \) and \( w \) are independent;
iv. The support of \( \theta \), \( \Theta \), satisfies Assumption 4 in Fox and Gandhi (2016);

then \( \beta_0, \beta_1 \), and the distribution of \( \theta \) are identified.

3.3. Bundles

Note that since I do not assume independence among \( \epsilon_y \) across \( y \), the multinomial choice model I study covers all bundles models (Gentzkow 2007; Dunker, Hoderlein, and Kaido 2017; Fox and Lazzati 2017). In particular, assume that there are \( J \) 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, \( \tilde{y} \in \tilde{Y} = \{0, 1\}^J \). For instance, \( \tilde{y} = (0, 1, 0, 1, 0, \ldots, 0) \) 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

\[
(\beta_0(w) + \beta_1(w)d + e) \sum_{j=1}^J \tilde{y}_j \tilde{z}_j + \tilde{e}_j,
\]

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

Corollary 3.4. Let \( J = 2 \) and

\[
\epsilon_{(1,0)} = \theta_1(w) + \epsilon_1, \quad \epsilon_{(0,1)} = \theta_2(w) + \epsilon_2,
\]

\[
\epsilon_{(1,1)} = \epsilon_{(1,0)} + \epsilon_{(0,1)} + \xi \theta_3(w),
\]

where \( \theta_i(\cdot), i = 1, 2, 3 \), are some unknown functions, and \( (\epsilon_1, \epsilon_2, \xi) \in \mathbb{R}^2 \times \mathbb{R}_+ \). Suppose

i. Assumptions of Propositions 3.1 and 3.2 or Proposition 3.3 hold;
ii. \( R_w = \mathbb{R} \) for all \( w \);
iii. \( (\epsilon_1, \epsilon_2)|w = w \) has an everywhere positive Lebesgue density on its support for all \( w \in W \);
iv. \( E[\xi_i|w = w] = 0 \) and \( E[\xi|w = w] = 1 \) for all \( w \in W \) and \( i = 1, 2, \)
then \( \theta_i(\cdot) \), \( i = 1, 2, 3 \), and the cdfs \( F_{\xi_i|w}, i = 1, 2, \) and \( F_{\xi|w} \) are identified.

### 4. Estimation of \( \beta \)

**Proposition 3.3** constructively identifies \( \beta_0 \) and \( \beta_1 \). 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 \( e \).\(^{10}\) Moreover, to simplify the exposition, I assume that there are no nonexcluded covariates \( w \) (i.e., \( \beta_0(\cdot) \) and \( \beta_1(\cdot) \) are constant functions). Note that, even though \( \beta_0 \) and \( \beta_1 \) are finite-dimensional parameters and the distribution of \( e \) is assumed to be known, the model is still semiparametric since the distribution of \( e \) is not parametric.

The first ingredient of the estimator is a nonparametric estimator of \( p_0(\cdot) = \Pr(y = 0|x = \cdot) \), \( p_0(\cdot) \). 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 (iid) observations of covariates and a binary random variable that indicates whether the product was purchased or not \( \{I(y^{(i)} = 0), x^{(i)}\}_{i=1}^n \), define

\[
\hat{p}_0(x) = \psi^K(x)^T (\Psi^T \Psi)^{-1} \sum_{i=1}^n \psi^K(x^{(i)}) I(y^{(i)} = 0),
\]

where \( \psi^K(\cdot) \) is a vector of orthonormal basis functions based on products of powers of components of \( x \), \( \Psi = (\psi^K(x^{(1)}), \psi^K(x^{(2)}), \ldots, \psi^K(x^{(m)}))^T \), and \( (\Psi^T \Psi)^{-1} \) is the Moore–Penrose generalized inverse. Assume that the sum of powers of components of \( x \) in \( \psi^K(\cdot) \) is monotonically increasing in \( K \).

The sign of \( \beta_1 \) can be trivially estimated from \( \hat{p}_0 \) since

\[
\text{sign}(\beta_1) = \text{sign}(p_0((d', z') - p_0((d, z)))) \text{sign}(z_{d'}) \text{sign}(d' - d)
\]

if \( z \geq 0 \) or \( z \leq 0 \) with \( z_{d'} \neq 0 \). Hence, for simplicity I assume that \( \beta_1 > 0 \).

The identification result in **Proposition 3.3** is constructive and provides a closed form expression for \( \beta \) as a functional of \( p_0 \) (see Appendix A.3). Given the nonparametric power series estimator \( \hat{p}_0 \), the plug-in estimator of \( \beta \) is

\[
\hat{\beta}_1 = \sqrt{\frac{\sum_{i=1}^n \hat{p}_{111} \left( x^{(i)} \right) \hat{p}_1 (x^{(i)}) - \hat{p}_{111} (x^{(i)})}{\sum_{i=1}^n \hat{p}_{12} \left( x^{(i)} \right) \hat{p}_1 (x^{(i)}) - \hat{p}_{2} (x^{(i)}) \hat{p}_{11} (x^{(i)}) - \hat{p}_{11} (x^{(i)})}}.
\]

where

\[
\hat{p}_1 = \frac{1}{\hat{\beta}_1} \left[ \sum_{i=1}^n \hat{p}_2 \left( x^{(i)} \right) - d^0 \hat{p}_1 (x^{(i)}) \right] - \frac{1}{\hat{\beta}_1} \sum_{i=1}^n \hat{p}_{111} (x^{(i)}) - \frac{1}{\hat{\beta}_1} \sum_{i=1}^n \hat{p}_1 (x^{(i)})
\]

\[
\hat{p}_0 = \hat{\beta}_1 \left[ \sum_{i=1}^n \hat{p}_2 \left( x^{(i)} \right) - d^0 \hat{p}_1 (x^{(i)}) \right] - \frac{1}{\hat{\beta}_1} \sum_{i=1}^n \hat{p}_{111} (x^{(i)}) - \frac{1}{\hat{\beta}_1} \sum_{i=1}^n \hat{p}_1 (x^{(i)})
\]

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

\[
\hat{v}_1 = -\left[ \sum_{j=1}^f \hat{p}_{1111}(x) f_j(x) + \sum_{j=1}^f \hat{p}_{111}(x) \hat{d}_j(x) f_j(x) + \sum_{j=1}^f \hat{p}_{11}(x) \hat{d}_{j,j}(x) f_j(x) + \sum_{j=1}^f \hat{p}_{1}(x) \hat{d}_{j,j,j}(x) f_j(x) \right] / f(x),
\]

\[
\hat{v}_2 = \left[ \beta_1 \left( 1 - f \right) f(x) + d \hat{d}_f(x) - \sum_{j=1}^f \hat{d}_{j,j} f(x) \right] / f(x),
\]

\[
\hat{v}(x) = (\hat{v}_1(x), \hat{v}_2(x)),
\]

where \( f(x) \) is the pdf of \( x \), and \( p_1, p_{11}, p_{111}, \) and \( p_{1111} \) are first, second, third, and fourth derivatives of \( p_0 \) with respect to \( d \), respectively.

**Assumption 5.**

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

ii. \( f(x) \) is bounded away from zero on the interior of \( X \).

iii. \( f(x), \hat{d}_f(x), \hat{d}_{j,f}(x), \) and \( \hat{d}_{j,j,f}(x) \) equal to zero at the boundary of \( X \) for all \( y \).

iv. \( E[\hat{v}(x) \hat{v}(x)^T] \) 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, Stock, and Stoker (1989), to achieve \( \sqrt{n} \)-consistency the estimator I need to impose restrictions on the behavior of \( f(x) \) on the boundary of its support. Since Powell, Stock, and Stoker (1989) work with the first derivative they only require \( f(x) \) 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 \( x \) (i.e., \( \log f(x) \)) 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

\[
G = \left( \begin{array}{cc}
2 \hat{p}_1 & 0 \\
0 & 0
\end{array} \right) \left( \begin{array}{cc}
\hat{E}[\hat{p}_{12}(x) \hat{p}_1(x) - \hat{p}_2(x) \hat{p}_{11}(x) - \hat{p}_1(x)^2] & 0 \\
0 & \hat{E}[\hat{p}_1(x)^2]
\end{array} \right)^{-1}.
\]
Proposition 4.1. If (a) \(\{I(y^{(i)} = 0), x^{(i)}\}_{i=1}^n\) are iid; (b) Assumptions 2, 4, and 5 are satisfied, and Assumption 4(iii) is satisfied for all \(x^{*x} = (d^{*x}, z^{*x}) \in X\); (c) \(K^0/n \to n \to \infty, 0\), then

\[
\sqrt{n} (\hat{\beta} - \beta) \to_d N(0, V),
\]

where \(V = GE[v(x)v(x)^T p_0(x)(1 - p_0(x))] G^T\).

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 \(\beta\) is estimated, one can construct a sieve maximum-likelihood estimator of \(F_{x}\) since

\[
Pr(y = 0|x = x) = \int_{\mathbb{R}} F_e(tz_1, tz_2, \ldots, tz_t) \phi(t + \beta_0 + \beta_1 d) \, dt
\]

where \(\phi(\cdot)\) is the standard normal pdf. Thus, one can find the maximizer of

\[
\max_{F_{x} \in \mathbb{R}} \sum_{i=1}^n \left[ I(y^{(i)} = 0) \log \left( \int_{\mathbb{R}} F(tz_1, tz_2, \ldots, tz_t) \phi(t + \beta_0 + \beta_1 d) \, dt \right) \right] + \left[ I(y^{(i)} = 0) \log \left( 1 - \int_{\mathbb{R}} F(tz_1, tz_2, \ldots, tz_t) \phi(t + \beta_0 + \beta_1 d) \, dt \right) \right],
\]

where \(\{F_{x}\}_{n=1}^\infty\) is a sequence of sieve spaces for \(F_{x}\). Inference on known functionals of \(\beta\) and \(F_{x}\) (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:

\[
y = I((\beta_0 + \beta_1 d + e)x + \beta_3 + e \geq 0),
\]

where \(\beta_0 = -0.5, \beta_1 = 1, \text{ and } e\) is a standard normal random variable. The random intercept \(\beta_3 + e\) is independent from \(x\) and \(e\) with mean \(\beta_3 = 0.5\). The observed covariates \(x = (d, z)\) are distributed according to a monotone transformation of a bivariate normal distribution: \(x = 5(\arctan(x)/\pi + 0.5)\), where \(x\) 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 \(x\) has bounded support.

I consider several data-generating processes (DGPs). The first one (DGP-0) is when \(e\) is a standard normal random variable. The next five DGPs correspond to \(e\) being an equally weighted mixture of three unit-variance normal distributions with mean \(-t, 0, t\), and for \(t \in \{1, 2, 3, 4, 5\}\) (DGP-\(t\)). For every \(t\) the distribution of \(e\) 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 \(e\).

Each experiment is conducted 1000 times for every DGP for three 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 \(\beta_1\) are presented in Table 1. As expected, the bias decreases with the sample size.\(^{14}\) However, there is not much variation across DGPs.\(^ {15}\)

6. Illustrative Empirical Application

To illustrate the empirical importance of the relaxation of the parametric assumptions about the distribution of \(F_{x}\) and the proposed estimation procedure, I analyze margarine purchasing decisions of households from Springfield, MO, USA, using the multinomial choice model with normally distributed \(e\). 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, Van Aelst, and Van den Poel (2016) focused on five brands instead of 10 and transformed this dataset to a cross-section with 242 households. In particular, every observation contains 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 five 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.

---

12 Both \(\beta\) and \(F_{x}\) can be estimated in one step by the sieve maximum-likelihood estimator. In this case, however, the estimator of \(\beta\) may not be \(\sqrt{n}\) consistent.

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

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.

17 Income and prices are measured in thousands of U.S. dollars and U.S. dollars, respectively.
the random intercept—multinomial logit. Formally, I estimate

First, I assume the most common parametric specification for

Parametric Estimation

Table 2. Summary statistics for products.

| Brand          | Share | Average price | Median price | Min price | Max 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

\[
\delta \mathbf{d} + (\beta_0 + \beta_1 \mathbf{d} + \mathbf{e}) \mathbf{p}_y + \tilde{\epsilon}_y.
\]

The random coefficient \( \delta \) 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 \( \tilde{\epsilon}_y \) (e.g., it is not assumed that it has zero mean).\(^{18}\)

This utility specification corresponds to the “preference shifter” specification in Griffith, Nesheim, and O’Connell (2018). There is no information about those who did not purchase any margarine products, thus, I analyze the choices of those who already 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

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

and the utility from purchasing Generic brand is 0. Hence, I can define \( \mathbf{z}_y = \mathbf{p}_y - \mathbf{p}_0 \) and \( \mathbf{e}_y = \tilde{\epsilon}_y - \tilde{\epsilon}_0, y = 1, 2, 3, 4 \), 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, \( \max_{\mathbf{z}_y}, \max_{\min_{\mathbf{z}_y}, \min_{\mathbf{z}_y}} = -0.15 \). At the same time, there is still variation in relative prices \( \mathbf{z}_y \) and income \( \mathbf{d} \). This variation allows me to recover \( \beta \) without specifying the distribution of \( \mathbf{e} \).

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).

Semiparametric Estimation

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

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

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

The estimates of \( \beta_0 \) and \( \beta_1 \) are \( \hat{\beta}_0 = -6331.94 \) (standard error \( = 17.19 \)) and \( \hat{\beta}_1 = -19.69 \) (standard error \( = 514.48 \)), respectively. As expected, the sign of \( \hat{\beta}_0 \) is negative. The coefficient in front of the income variable, \( \hat{\beta}_1 \), is negative and not significant at the 5% significance level. Although income does not matter much, the overall sensitivity to prices (mostly captured by \( \hat{\beta}_0 \) 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.\(^{19}\)

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:

\[
\mathbb{I}(y \neq 0) \{y_y + (\beta_0 + \beta_1 \mathbf{d} + \mathbf{e}) \mathbf{z}_y\} + \alpha \mathbf{e}_y,
\]

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

The estimates of \( \beta_0 \) and \( \beta_1 \) 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% 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 \( \{\mathbf{d}^{	op}\hat{\beta}_0/\hat{\beta}_1\} \approx 0.055, \) 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 \( \hat{\beta}_0 \) 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 logit estimator \( \hat{\beta} \) and my semiparametric estimator \( \hat{\beta} \) is substantial (e.g., \( \hat{\beta}_1/\hat{\beta}_1 > 10^3 \) ). 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.\(^{21}\)

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

\(^{19}\)For example, in U.K. households spend about 1% of their grocery expenditures on margarine and butter (Griffith, Nesheim, and O’Connell 2018).

\(^{20}\)Use the tensor product of the fourth degree Chebyshev polynomials for \( \mathbf{d} \) and the first degree Chebyshev polynomials for every \( \mathbf{z}_y \).

\(^{21}\)This empirical finding is in line with the simulation results, presented in Appendix B.
Appendix 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 y from a known finite set Y, a vector of observed exogenous characteristics x ∈ X ⊆ R^d, d < ∞, that can be partitioned into x = (d, z), and a vector of unobserved indexes s ∈ S ⊆ R^d_i.

Assumption 6 (Data). There exists Y^* ⊆ Y such that the analyst observes (can consistently estimate) μ(y|x) = Pr(y = y|x = x) for all x ∈ x and y ∈ Y^*.

Assumption 7. There exists h_0 : Y^* × S → [0, 1], such that Pr(y = y|x = x, s = s) = h_0(y,s), for all y ∈ Y^*, x ∈ X, and s ∈ S.

Assumption 8 is an outcome restriction that requires d and z to affect distribution over outcomes in Y^* only via the distribution of s.

Assumption 8 (Bounded completeness). There exists X' ⊆ X such that the family of distributions \{F_{y|x}|x ∈ X'\} is boundedly complete.

Proposition A.1. Under Assumptions 6–8, h_0 is identified from μ up to F_{y|x}.

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

\[ \forall x \in X : μ(y|x) = \int_S h(y,s)dF_{y|x}(s|x). \]

Suppose that there exists h with h(y,s) ≠ h_0(y,s) for all s in some nonzero-measure set S' such that

\[ \forall x \in X : μ(y|x) = \int_S h(y,s)dF_{y|x}(s|x) = \int_S h_0(y,s)dF_{y|x}(s|x). \]

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

A.1. Nonparametric Identification

Given a collection of random variables \{ξ_i\}_{i=1,...,d} d < ∞, I say that ξ_i is redundant if there exists j ≠ i such that ξ_i = ξ_j a.s. Nonredundant elements of \{ξ_i\}_{i=1,...,d} is the largest subset of \{ξ_i\}_{i=1,...,d} such that none of its elements are redundant.

Assumption 9. i. The latent s = (s_i)_{i=1,...,d} satisfies

s_i = ξ_i(β_{0,i} + β_{1,i}d_i + e_i) a.s.

where β_{0,i} and β_{1,i} are some unknown parameters such that β_{1,i} ≠ 0 for all i = 1, …, d_i;

ii. Nonredundant elements of \{ξ_i\}_{i=1,...,d} are mean-zero and variance-one independent random variables that are independent of x;

iii. For some y^* ∈ Y^*, h_0(y^*,·) has bounded derivatives up to order κ and \frac{∂^κ}{∂z^κ} h_0(y^*,z) |_{z=0} ≠ 0 for all l ≤ κ and all i = 1, …, d_i;

iv. The support of x, which consists of nonredundant elements of \{d_i\}_{i=1,...,d} and all of \{ξ_i\}_{i=1,...,d}, contains x^* with an open neighborhood such that x^* ⊆ 0 for all i = 1, …, d_i;

v. The sign of either β_{0,i} or β_{1,i} is known for every i = 1, …, d_i.

Let β_0 = (β_{0,i})_{i=1}^{d_i} and β_1 = (β_{1,i})_{i=1}^{d_i}.

Proposition A.2. If Assumptions 6, 7, and 9 hold, then β_0, β_1, and E[ξ^2], i = 1, …, d_i, 0 ≤ l ≤ κ, are identified.

Proof. Given a family x = (x_i)_{i=1}^{n} and a particular index value k ∈ K, let x_k denote (x_i)_{i=1}^{n} and set z_i = 0, y^* in Y^* from Assumption 9iii. To simplify notation, let F_0 : R → R and η : R^2 → R such that F_0(t) = h_0(y^*(0,0, …, 0)), where the only nonzero component in the second argument of h_0 is the ith component, and η(ds,zi) = μ(y^*(x)). Note that Assumption 9iii together with the dominated convergence theorem imply that F_0 has bounded derivatives up to order κ.

Assumptions 7 implies that

\[ η(ds,zi) = \int F_0((β_{0,i} + β_{1,i}d_i + e_i)zi)dF_{ξ|x}(x). \]

Next, since ξ_i and x are independent and h_0(y^*,·) is κ-times differentiable with bounded derivatives, the dominated convergence theorem implies that (I dropped the subscript i from the notation)

\[ \frac{∂^l}{∂z^l} η(ds,zi) = \frac{∂^l}{∂z^l} F_0((β_0 + β_1d + e)z)dF_{ξ|x}(x) \]

for any l ≤ κ. Hence, since derivatives of h_0(y^*,·) are bounded, applying the dominated convergence theorem again I get that

\[ \lim_{z_i → 0} \frac{∂^l}{∂z^l} η(ds,zi) = β_1^l \int \frac{∂^l}{∂z^l} F_0(0)dF_{ξ|x}(x). \]

and, thus, β_1^l F_0(0) is identified for any l ≤ κ. Similarly note that, since h_0(y^*,·) has bounded derivatives,

\[ \frac{∂^l}{∂z^l} F_0(0) = \int \frac{∂^l}{∂z^l} F_0(0)dF_{ξ|x}(x) \]

(1)

for every l ≤ κ. Hence, since E[ξ_i] = 0 and β_0 β_i F_0(0) is identified, β_0 β_i F_0(0) = \frac{∂^l}{∂z^l} F_0(0) d_i β_i F_0(0) d_i 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 E[ξ_i] = 0 and E[ξ_i^2] = 1, we also can derive that

\[ \frac{∂^2}{∂z^2} η(ds,zi) = \int \frac{∂^2}{∂z^2} F_0(0)(β_0 + β_1d + e)2dF_{ξ|x}(x) \]

\[ = 0^2 F_0(0) \left[ (β_0 + β_1d)^2 + 1 \right]. \]

Hence, \frac{∂^2}{∂z^2} η(ds,zi) = β_1^2 F_0(0) \left[ (β_0 + β_1d)^2 + 1/β_1^2 \right]. As a result, since we identified β_0/β_1 and β_1^2 F_0(0) in the previous steps,

\[ 1/β_1^2 = \frac{∂^2}{∂z^2} η(ds,zi) - (β_0/β_1 + d)^2 \]

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 \frac{∂^2}{∂z^2} F_0(0) for all l ≤ κ.
To identify all moments of \( \mathbf{e} \) up to order \( \kappa \), I use Equation (1) to derive the following recursive equation

\[
E[\mathbf{e}^k] = \frac{\partial^k}{\partial t^k} F(0) - \sum_{k=1}^{l} (\partial \beta_0 + d)^k E[\mathbf{e}^{k-l}].
\]

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

Note that Proposition A.2 allows \([ z] \) and \( d \), and nonredundant elements of \( \{ e \} \) and \( \{ d \} \) to have different cardinality. If the cardinality of nonredundant elements of \( \{ e \} \) and \( \{ d \} \) is the same, then the assumption that \( \{ e \} \) and \( \{ d \} \) are independent can be relaxed. In this case, using a similar strategy, one can identify recursively \( E[\mathbf{e}^k] \) for all positive \( i \), and set of nonnegative elements \( \{ e \} \) such that \( \sum_{i=1}^{\kappa} k_i \leq \kappa \). For instance, if \( d_4 = 2 \), then for \( E(\mathbf{e}) = h(y^*, \mathbf{v}) \) I have that

\[
\eta(d, z) = \int_{R^2} F((\beta_0, \beta_{1,i} d_1 + e_1)z_1, \beta_0 + \beta_{1,i} d_2 + e_2)z_2) d\mathbf{e}(e).
\]

Thus, given that \( \beta_0 \) and \( \beta_{1,i} \) are identified, we can identify,

\[
\lim_{|z|\to 0} \frac{\partial^2_{d_1, d_2} \eta(d, z)}{d_1 d_2} = \beta_{1,1} \beta_{1,2} \int_0^\infty \frac{\partial^2_{d_1, d_2} F(0) d\mathbf{e}(e)}{d_1 d_2}.
\]

As a result, the partial derivative with respect to \( \mathbf{e} \) and \( \mathbf{e}_i \)

\[
\frac{\partial^2_{d_1, d_2} \eta(d, 0)}{d_1 d_2} = \beta_{1,1} \beta_{1,2} \int_0^\infty \frac{\partial^2_{d_1, d_2} F(0) d\mathbf{e}(e)}{d_1 d_2}.
\]

Normal Random Coefficient

Assumption 10. i. The latent \( s = (s_i)_{i=1, \ldots, d} \) satisfies

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

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

ii. \( \{ e_i \} \) are iid standard normal random variables that are independent of \( s \);

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 \( i = 1, \ldots, d \).

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.

Assumption 11. For every \( i = 1, 2, \ldots, d \), 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.

Proof. Note that \( h_0 \) is identified up to \( \beta_0 \) and \( \beta_1 \) because of completeness of the family of normal distributions and Proposition A.1. Hence, I only need to show that \( \beta_0 \) and \( \beta_1 \) are identified. Fix some \( i \in \{ 1, 2, \ldots, d \} \), \( z \), and \( d_i \) in the support. Take \( y^* \) from Assumption 11. To simplify notation, let \( F_0 : R \to R \) and \( \eta : R^2 \to R \) be functions such that

\[
F_0(s) = \int_{R} h_0(y^*, \eta) \phi(z_k / z_k - \beta_0 d_k) z_k d\mathbf{z},
\]

where \( \phi(\cdot) \) is the standard normal pdf and \( \eta(d, z) = \mu(y^*)d, \eta \). Assumptions 7 and 10 imply that \( \eta(d, z) = \int_{R} h_0(s) \phi(s / z - \beta_0 d) ds \).

Next, note that since \( \hat{\beta}_0^2 \phi(x) = -\phi(x) - x_0 \phi(x) \) the following system of equations holds

\[
\begin{align*}
\hat{\beta}_0 \hat{\beta}_1 &= \int_{R} F_0(t) \partial_\mathbf{d} \partial_\mathbf{e} \phi(t / z - \beta_0 d) dt, \\
\hat{\beta}_2 \hat{\beta}_1 &= \int_{R} F_0(t) \partial_\mathbf{d}^2 \partial_\mathbf{e} \phi(t / z - \beta_0 d) dt \\
&= -\hat{\beta}_1^2 \hat{\beta}_1 \eta(d, z) - \beta_1 (\beta_0 + \beta_1 d) \partial_\mathbf{d} \eta(d, z) \\
&= \hat{\beta}_1^2 \int_{R} F_0(t) \partial_\mathbf{d} \phi(t / z - \beta_0 d) dt / z.
\end{align*}
\]

Moreover, \( \partial_\mathbf{d} \eta(d, z) = -\int_{R} F_0(t) \partial_\mathbf{d} \phi(t / z - \beta_0 d) dt / z^2 \). Hence, \( \hat{\beta}_2 \hat{\beta}_1 \eta(d, z) = -\hat{\beta}_1^2 \eta(d, z) - \beta_1 (\beta_0 + \beta_1 d) \partial_\mathbf{d} \eta(d, z) + \hat{\beta}_1^2 \partial_\mathbf{d} \eta(d, z) \).

Equivalently,

\[
\begin{align*}
\beta_0 &= z \hat{\beta}_1 \eta(d, z) - \hat{\beta}_1 \eta(d, z) - d - \hat{\beta}_1^2 \eta(d, z) / \hat{\beta}_1^2;
\end{align*}
\]

Replacing \( \eta(d, z) \) by \( z_\eta(d, z) \), I get

\[
\begin{align*}
\beta_0 &= z \hat{\beta}_1 \eta(d, z) - d \hat{\beta}_1 \eta(d, z) - \hat{\beta}_1^2 \eta(d, z) / \hat{\beta}_1^2.
\end{align*}
\]

Thus, \( \beta_0 / \hat{\beta}_1 \) is identified up to \( \hat{\beta}_1 \). Differentiating the last equation with respect to \( d \) leads to the following equation

\[
\begin{align*}
\frac{1}{\hat{\beta}_1} &= \partial_\mathbf{d} \left[ \frac{z \hat{\beta}_1 \eta(d, z) - d \hat{\beta}_1 \eta(d, z) - \hat{\beta}_1^2 \eta(d, z) / \hat{\beta}_1^2}{\hat{\beta}_1 \eta(d, z)} \right] / \hat{\beta}_1 \left[ \frac{\partial^2_{\mathbf{d}^2} \eta(d, z)}{\hat{\beta}_1 \eta(d, z)} \right];
\end{align*}
\]

Hence, if

\[
\partial_\mathbf{d} \left[ \frac{\partial^2_{\mathbf{d}^2} \eta(d, z)}{\hat{\beta}_1 \eta(d, z)} \right] \neq 0
\]

for some \( d \) and \( z \), then \( \hat{\beta}_1 \) is identified. Suppose this is not the case. That is, for all \( d \) and \( z \) \( \partial_\mathbf{d} \left[ \frac{\partial^2_{\mathbf{d}^2} \eta(d, z)}{\hat{\beta}_1 \eta(d, z)} \right] = 0 \). Equivalently, \( \hat{\beta}_1^2 \left[ \frac{\partial^2_{\mathbf{d}^2} \eta(d, z)}{\hat{\beta}_1 \eta(d, z)} \right] \neq 0 \) for all \( d \) and \( z \). The latter would imply that either \( \eta(d, z) = K_1(z)K_2(z) + K_2(z) \) or \( \eta(d, z) = K_1(z) \) for some 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, 1

(22) I can differentiate under the integral sign since (i) \( h_0 \) being bounded implies \( F_0 \) is bounded, (ii) all derivatives of the standard normal pdf are bounded.
can conclude that for some $d$ and $z$ Equation (5) is satisfied. Thus, $\beta_1^2$ is identified (hence, $|\beta_1|$ is also identified). Hence, I identify $\beta_0/\beta_1$. If $\beta_0/\beta_1 = 0$, then the sign of $\beta_0$ is identified from Assumption 10(iv). If $\beta_0/\beta_1 \neq 0$, then the sign of either $\beta_1$ or $\beta_0$ is identified from Assumption 10(iv). Knowing the sign of, say, $\beta_0$ and $\beta_0/\beta_1$ identifies $\beta_1$ and $\beta_0$. Going back to the original notation I identify $\beta_1$ and $\beta_0$. The conclusion of the proposition then follows from the fact that the choice of $i$ was arbitrary.

Note that for identification of $\beta_1$ and $\beta_0$, I do not need to exclude all polynomial functions of $d$, since instead of differenitizing 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 $\eta(d, z) = K_1(z)K_2d + K_3(z)$ 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 3.3

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 article. Fix some arbitrary $w \in W$. To prove Propositions 3.1 and 3.3(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 $h(0, s) = F_{e|w}(0, \ldots, 0, 0)$, where the only nonzero component corresponds to $\hat{y}$ by Assumption 3(ii). To show validity of Assumption 7 in Proposition A.3, note that Assumption 3(iii) or Assumption 4(ii) there exist $z^*$ and $[\lambda_j]_{j=1}^l$ with some open neighborhood such that $z_j^* = \lambda_j z_j^*$ for all $y \in Y$ with $\min_j \lambda_j > 0$. Note that e and z are independent conditional on w, I have that for $x^* = (d^*, z^*, w)$

$$
\mu(0|x^*) = \int_R F_{e|w}(-z^*_1(\beta_0(w) + \beta_1(w)d^*) + c), \ldots, -\lambda_l z^*_l(\beta_0(w) + \beta_1(w)d^*) + c)dwF_{e|w}(e|w).
$$

Hence, Assumptions 7 is satisfied for $h(0, s) = F_{e|w}(s, \lambda_j z^*_j)$.

### 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 $0 \leq p_0(v) \leq 1$ for all $v$ and

$$
p_0(x) = \int_R h_0((\beta_0 + \beta_1 d + e)\phi(e)de = \int_R h_0(v\phi(v/z_1 - \beta_1 d - \beta_0)dv/z_1.
$$

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 $x = x$ and $e = e$ equals to 0 for all $e$. Since $\xi_1$ is independent of $x$ and $e$, I have that $\xi_1 \geq -z_1(\beta_1 + \beta_1 d + e)$ with probability 1 for all $e$, which is not possible since $e$ 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 $p_1(x) = \partial p_0(x)$. Since $\partial \phi(t) = -t\phi(t)$,

$$
|p_1(x)| = \left|\beta_1 \int_R h_0(v)/(v/z_1 - \beta_1 d - \beta_0)\phi(v/z_1 - \beta_1 d - \beta_0)dv/z_1
\right| = \left|\beta_1 \int_R h_0((\beta_0 + \beta_1 d + c)z1)\phi(c)(\beta_0 + \beta_1 d + c)\phi(c).\right|
$$

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

$$
p_2(x) \leq |p_0(x)| + \int_R h_0((\beta_0 + \beta_1 d + e)\phi(c)(\beta_0 + \beta_1 d + c)\phi(c).
$$

Hence, given bounded support of $x$, I can conclude that sup $|p_2(x)|$ 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 3.3 we used derivatives of $\eta(d, z_1)$ to identify $\beta_1$. In particular, we can take $\eta(d^{**, z^{**}}) = \mu(0|x^*)$, where $x^* = (d^{**}, z^{**})$, and $\lambda_2 = z_2^*/z_1^*$. As a result, $\partial_{d_1} \eta(d^{**, z^{**}}) = \sum_j \partial_{d_1} \eta_j(0|x^*)$. Since $\lambda_2 = z_2^*/z_1^*$, I get that $\partial_{d_1} \eta_j(0|x^*) = \sum_j \eta_j \partial_{d_1} \eta_j(0|x^*)$. Hence, if Assumption 4(iii) is satisfied not just for one $(d^{**}, z^{**})$ but for all, then for all $x$

$$
\beta_1^1 = \frac{\partial^2_{d_1} p_0(x)\partial_{d_1} p_0(x) - [\partial_{d_1} p_0(x)]^2}{\partial^2_{d_1} p_0(x)\partial_{d_1} p_0(x) - [\partial_{d_1} p_0(x)]^2},
$$

$$
\beta_0^1 = \frac{\sum_j \partial_{d_1} \eta_j(0|x^*) - \partial_{d_1} \eta_j(0|x^*)}{\partial_{d_1} \eta_j(0|x^*) - \partial_{d_1} \eta_j(0|x^*)}.\beta^1_1 - \frac{\partial^2_{d_1} p_0(x)1}{\partial_{d_1} p_0(x)\beta^1_1}.
$$

**Step 3.** Combining the bounds for the derivatives from Step 1, the uniform weak law of large numbers, and consistency of $p_0$, I can deduce that

$$
\frac{1}{n} \sum_{i=1}^n \tilde{p}_{11} \left( x_i \right) - \tilde{p}_{11} \left( x \right)^2 \rightarrow_p \mathbb{E} \left[ p_{11}(x)p_1(x) - p_{11}(x)^2 \right],
$$

$$
\frac{1}{n} \sum_{i=1}^n \tilde{p}_{12} \left( x_i \right) - \tilde{p}_{12} \left( x \right) - \tilde{p}_{11}(x) \tilde{p}_{11}(x) - \tilde{p}_{11}(x) \left( x \right)^2 \rightarrow_p \mathbb{E} \left[ p_{12}(x)p_1(x) - p_{12}(x)p_1(x) \right],
$$

$$
\frac{1}{n} \sum_{i=1}^n \tilde{p}_{12} \left( x_i \right) - \tilde{p}_{12} \left( x \right) - \tilde{p}_{11}(x) \tilde{p}_{11}(x) - \tilde{p}_{11}(x) \left( x \right)^2 \rightarrow_p \mathbb{E} \left[ p_{12}(x)p_1(x) - p_{12}(x)p_1(x) \right].
$$
\[ \frac{1}{n} \sum_{i=1}^{n} \beta_2(x^{(i)}) - d^{(i)} \hat{p}_1(x^{(i)}) \rightarrow_p E[p_2(x) - dp_1(x)], \]
\[ \frac{1}{n} \sum_{i=1}^{n} \beta_{11}(x^{(i)}) \rightarrow_p E[p_{11}(x)], \quad \frac{1}{n} \sum_{i=1}^{n} \hat{p}_1(x^{(i)}) \rightarrow_p E[p_1(x)]. \]

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

**Step 4.** Consider

\[ G_n = \frac{1}{n} \sum_{i=1}^{n} \left( \left( \hat{p}_{11}(x^{(i)}) \hat{p}_1(x^{(i)}) - \hat{p}_{11}(x^{(i)}) \right)^2 - \hat{p}_{11}(x^{(i)}) \right). \]

To prove asymptotic normality of \( G_n \), I will use Theorem 6 in Newey (1997). The data is assumed to be iid, 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 \( a(p_0) = (a_1(p_0), a_2(p_0)) \) with

\[ a_1(p_0) = E[p_{111}(x)p_1(x) - p_{111}(x)^2], \]
\[ a_2(p_0) = E[p_2^2(x) - dp_1(x)] - p_{111}(x). \]

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

\[ D_1(g_0) = E[p_{111}(x)g_1(x) + g_{111}(x)p_1(x) - 2p_{11}(x)g_{11}(x)], \]
\[ D_2(g_0) = E[p_2^2 g_2(x) - d g_1(x)] - g_{11}(x). \]

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

\[ E[p_{111}(x)g_1(x)] = -E[\partial x_1[p_{111}(x)f_x(x)]g_0(x)/f_x(x)], \]
\[ E[p_1(x)g_{111}(x)] = -E[\partial x_1^2[p_1(x)f_x(x)]g_0(x)/f_x(x)], \]
\[ E[p_{11}(x)g_{11}(x)] = E[\partial x_1^2[p_{11}(x)f_x(x)]g_0(x)/f_x(x)], \]
\[ E[d g_1(x)] = -E[\{f_x(x) + d f_x(x)g_0(x)/f_x(x) \}], \]
\[ E[g_{11}(x)] = E[\partial x_1^2 f_x(x)g_0(x)/f_x(x)], \]
\[ E[g_{22}(x)] = -E[\{f_x(x) + f_x(x) \partial x_1 f_2(x)g_0(x)/f_x(x) \}]. \]

As a result,

\[ D_1(g_0) = -E[\{4p_{1111}(x)f_x(x) + 8p_{111}(x)\partial x_1 f_x(x) + 5p_{11}(x)\partial x_1^2 f_x(x) + p_1(x)\partial x_1^3 f_x(x)]g_0(x)/f_x(x)], \]
\[ D_2(g_0) = E[p_2^2(d f_x(x) - z_1 \partial z_2 f_x(x)) - \partial x_1^2 f_x(x)]g_0(x)/f_x(x). \]

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

\[ \tilde{G} = E[p_{111}(x)p_1(x) - p_{111}(x)^2] \]
\[ \tilde{\Sigma} = E[p_2^2(x) - dp_1(x)] - p_{111}(x) \]
\[ \tilde{\gamma} = \psi \tilde{G} \tilde{\gamma} \]
\[ \hat{\gamma} = \psi \tilde{G} \hat{\gamma} \]
\[ \hat{\beta} = \psi \tilde{G} \hat{\beta} \]
\[ \tilde{\psi} = \psi \tilde{G} \tilde{\psi} \]
\[ \tilde{\psi} = \psi \tilde{G} \tilde{\psi} \]

and \( \tilde{V} = E[\tilde{v}(x)\tilde{v}(x)^T(p_0(x)(1 - p_0(x))]. \) Moreover, I can construct a consistent estimator of \( \tilde{V} \) using Theorem 6 in Newey (1997). In particular, let \( \tilde{\alpha}(p_0) \) be a sample counterpart of \( \alpha(p_0) \)

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

\[ \sqrt{n}(\hat{\beta} - \beta) \rightarrow_d \left( \begin{array}{c} 2\beta_1 \\ 0 \\ 1 \end{array} \right) \]
\[ \left( E[p_{12}(x)p_1(x) - p_2(x)p_{11}(x) - p_1(x)^2] \beta_1 E[p_1(x)] \right)^{-1} \]
\[ N(0, \tilde{V}). \]

**Step 5. Consistency of \( \tilde{V} = \tilde{G} \tilde{V} \tilde{G}^T \), where \( \tilde{G} = \left( \begin{array}{c} 2\beta_1 \\ 0 \\ 1 \end{array} \right) \)
\[ \tilde{\psi} = \left( n^{-1} \sum_{i=1}^{n} \hat{p}_{11}(x^{(i)}) \hat{p}_1(x^{(i)}) - \hat{p}_{11}(x^{(i)}) \hat{p}_1(x^{(i)}) - p_1(x)^2 \right)^{-1} \]
\[ \tilde{\psi} = \left( n^{-1} \sum_{i=1}^{n} \hat{p}_{11}(x^{(i)}) \hat{p}_1(x^{(i)}) - \hat{p}_{11}(x^{(i)}) \hat{p}_1(x^{(i)}) - p_1(x)^2 \right)^{-1} \]

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

**Appendix B: Additional Simulations**

Table 3 contains results for the mean absolute deviation (MAD) of my estimator of \( \beta_1 \). Similar to the bias, the MAD decreases with \( n \) and is of the same magnitude across DGPs.

Next, I estimated \( \beta_1 \) using two maximum-likelihood estimators. The first one (Probit) is based on the assumption that \( \epsilon \) is normal. The second one (Logit) is assumed to be logistic. 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 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  |
| 10,000          | 0.38  | 0.42  | 0.67  | 0.90  | 1.13  | 1.21  | 0.52  |

**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  | 8.74  | 17.01 | 0.59  |

Overall, the Logit estimator outperforms the Probit estimator for all DGP-0 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 DGP, 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.

Supplementary Materials

The supplementary materials contain the replication package for all simulation and estimation results.

Acknowledgments

I thank the editor and two anonymous referees for comments and suggestions that have greatly improved the article. I also thank Roy Allen, Victor H. Aguian, Tim Conley, Nirav Mehta, Salvador Navarro, Joris Pinkse, David Rivers, and Bruno Salcedo for their useful comments and discussions.

References

Allen, R., and Rehbeck, J. (2020), "Identification of Random Coefficient Latent Utility Models," arXiv preprint arXiv:2003.00276. [697,698]
Allenby, G. M., and Rossi, P. E. (1991), "Quality Perceptions and Asymmetric Switching Between Brands," *Marketing Science*, 10, 185–204. [700]
Andrews, D. W. K. (2011), "Examples of L2-complete and Boundedly-Complete Distributions," discussion paper 1801, Cowles Foundation. [695]
Bajari, P., Hong, H., and Ryan, S. P. (2010), "Identification and Estimation of a Discrete Game of Complete Information," *Econometrica*, 78, 1529–1568. [696]
Benoit, D. F., Van Aelst, S., and Van den Poel, D. (2016), "Outlier-robust Bayesian Multinomial Choice Modeling," *Journal of Applied Econometrics*, 31, 1445–1466. [700]
Berry, S., Levinsohn, J., and Pakes, A. (1995), "Automobile Prices in Market Equilibrium," *Econometrica: Journal of the Econometric Society*, 63, 841–890. [696,697]
Bajari, J., et al. (2020), "Identification and Estimation of Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112, 68–105. [696]
Berry, S. T. (1994), "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 25, 242–262. [697]
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. [696,697]
Bajari, J., et al. (2014), "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797. [697]
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. [704]
Canay, I. A., Santos, A., and Shaikh, A. M. (2013), "On the Testability of Identification in Some Nonparametric Models with Endogeneity," *Econometrica*, 81, 2535–2559. [697]
Chen, S., Khan, S., and Tang, X. (2016), "Informational Content of Special Regressors in Heteroskedastic Binary Response Models," *Journal of Econometrics*, 193, 162–182. [696]
Chen, X., and Liao, Z. (2014), "Sieve m Inference on Irregular Parameters," *Journal of Econometrics*, 182, 70–86. [700]
Chernozhukov, V., and Hansen, C. (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. [695]
Chernozhukov, V., Imbens, G. W., and Newey, W. K. (2007), "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139, 4–14. [695]
Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011), "Nonparametric Instrumental Regression," *Econometrica*, 79, 1541–1565. [695]
d’Haultfoeuille, X. (2011), "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 27, 460–471. [695]
Dunker, F., Hoderlein, S., and Kaido, H. (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. [696,698]
Fox, J. T. (2020), "A Note on Nonparametric Identification of Distributions of Random Coefficients in Multinomial Choice Models," Technical report. [696,697]
Fox, J. T., and Gandhi, A. (2016), "Nonparametric Identification and Estimation of Random Coefficients in Multinomial Choice Models," *The RAND Journal of Economics*, 47, 118–139. [696,697,698]
Fox, J. T., Il Kim, K., Ryan, S. P., and Bajari, P. (2012), "The Random Coefficients Logit Model is Identified," *Journal of Econometrics*, 166, 204–212. [696,697,698]
Fox, J. T., and Lazzati, N. (2017), "A Note on Identification of Discrete Choice Models for Bundles and Binary Games," *Quantitative Economics*, 8, 1021–1036. [696,698]
Fox, J. T., Yang, C., and Hsu, D. H. (2018), "Unobserved Heterogeneity in Matching Games," *Journal of Political Economy*, 126, 1339–1373. [696]
Gautier, E., and Hoderlein, S. (2015), "A Triangular Treatment Effect Model with Random Coefficients in the Selection Equation," arXiv preprint arXiv:1109.0362. [696,697]
Gautier, E., and Kitamura, Y. (2013), "Nonparametric Estimation in Random Coefficients Binary Choice Models," *Econometrica*, 81, 581–607. [696]
Gentzkow, M. (2007), "Valuing New Goods in a Model with Complementarity," Online Newspapers, *American Economic Review*, 97, 713–744. [698]
Gentzkow, M. (2007), "Valuing New Goods in a Model with Complementarity," *American Economic Review*, 97, 713–744. [698]
Griffith, R., Nesheim, L., and O’Connell, M. (2018), "Income Effects and the Welfare Consequences of Tax in Differentiated Product Oligopoly," *Quantitative Economics*, 9, 305–341. [701]
Heckman, J. (1990), "Varieties of Selection Bias," *The American Economic Review*, 80, 313–318. [696]
Hu, Y., and Schennach, S. M. (2008), "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216. [695]
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, 269–295. [696]
Kashaev, N., and Salcedo, B. (2021), "Discerning Solution Concepts for Discrete Games," *Journal of Business & Economic Statistics*, 39, 1001–1014. [696]
Kleiber, C., and Stoyanov, J. (2013), "Multivariate Distributions and the Moment Problem," *Journal of Multivariate Analysis*, 113, 7–18. [697]
Kline, B. (2016), "The Empirical Content of Games with Bounded Regressors," *Quantitative Economics*, 7, 37–81. [696]
Kleiber, C., and Stoyanov, J. (2013), "Multivariate Distributions and the Moment Problem," *Journal of Multivariate Analysis*, 113, 7–18. [697]
Lewbel, A. (1998), "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121. [696]

Lewbel, A., Yan, J., and Zhou, Y. (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, 76–104. [696]

Manski, C. F. (1985), “Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator,” Journal of Econometrics, 27, 313–333. [696]

Manski, C. F. (1988), “Identification of Binary Response Models,” Journal of the American Statistical Association, 83, 729–738. [696]

Mattner, L. (1993), “Some Incomplete but Boundedly Complete Location Families,” The Annals of Statistics, 21, 2158–2162. [695]

Matzkin, R. L. (1992), “Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models,” Econometrica, 60, 239–270. [696]

Matzkin, R. L. (2007), “Heterogeneous Choice,” Econometric Society Monographs, 43, 75. [696]

Nevo, A. (2000), “A Practitioner’s Guide to Estimation of Random-Coefficients Logit Models of Demand,” Journal of Economics & Management Strategy, 9, 513–548. [696,698,701]

Nevo, A. (2001), “Measuring Market Power in the Ready-to-Eat Cereal Industry,” Econometrica, 69, 307–342. [696,698,701]

Newey, W. K. (1994), “The Asymptotic Variance of Semiparametric Estimators,” Econometrica, 62, 1349–1382. [696]

——— (1997), “Convergence Rates and Asymptotic Normality for Series Estimators,” Journal of Econometrics, 79, 147–168. [699,700,705]

Newey, W. K., and Powell, J. L. (2003), “Instrumental Variable Estimation of Nonparametric Models,” Econometrica, 71, 1565–1578. [695,697]

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), “Semiparametric Estimation of Index Coefficients,” Econometrica, 57, 1403–1430. [695,699]

Shen, X., and Shi, J. (2005), “Sieve Likelihood Ratio Inference on General Parameter Space,” Science in China Series A: Mathematics, 48, 67–78. [700]

Tamer, E. (2003), “Incomplete Simultaneous Discrete Response Model with Multiple Equilibria,” The Review of Economic Studies, 70, 147–165. [696]

Thompson, T. S. (1989), “Identification of Semiparametric Discrete Choice Models,” University of Minnesota Centre for Economic Research discussion paper, no. 249. [696]