Semiparametric Identification and Estimation of Multinomial Discrete Choice Models using Error Symmetry*

Arthur Lewbel†     Jin Yan‡     Yu Zhou§

Original February 2019, revised February 2021

Abstract

We provide a new method to point identify and estimate cross-sectional multinomial choice models, using conditional error symmetry. Our model nests common random coefficient specifications (without having to specify which regressors have random coefficients), and more generally allows for arbitrary heteroskedasticity on most regressors, unknown error distribution, and does not require a “large support” (such as identification at infinity) assumption. We propose an estimator that minimizes the squared differences of the estimated error density at pairs of symmetric points about the origin. Our estimator is root N consistent and asymptotically normal, making statistical inference straightforward.

Keywords: Central Symmetry, Exclusion Restriction, Multinomial Discrete Choice

JEL Code: C14, C35

*We thank Songnian Chen, Jeremy Fox, Bruce Hansen, Shakeeb Khan, Brendan Kline, Rosa Matzkin, Byoung Park, Jack Porter, Andres Santos, Xiaoxia Shi, Matthew Shum, Christopher Taber, Yuanyuan Wan and Hanghui Zhang for valuable comments and discussions and seminar participants at Xiamen University, Academia Sinica, Renmin University of China, the 2019 Asia Meeting, China Meeting, and Australasian Meeting of the Econometric Society, and the 29th Annual Meeting of the Midwest Econometrics Group. Funding support from the DVPS of the Chinese University of Hong Kong is greatly appreciated. All errors are ours.

†Department of Economics, Boston College. E-mail: lewbel@bc.edu.
‡Department of Economics, The Chinese University of Hong Kong, Hong Kong. E-mail: jyan@cuhk.edu.hk.
§Department of Economics, School of Economics, Fudan University; Shanghai; email: yuzhou@fudan.edu.cn
1 Introduction

Traditional multinomial (sometimes also called polychotomous) choice models, such as multinomial logit (MNL) and multinomial probit (MNP), e.g., McFadden (1974), assume homoskedastic errors. However, real data often have substantial unobserved heterogeneity. See, e.g., Heckman (2001). We provide a new method to point identify preference parameters in cross-sectional multinomial choice models in the presence of general unobserved individual heterogeneity. Our identification is semiparametric, in that we do not specify the joint distribution of the latent errors, and we allow for arbitrary heteroskedasticity with respect to most regressors, including possible random coefficients. We also propose a corresponding estimator, and show that it is root $N$ consistent and asymptotically normal.

Many empirically popular multinomial choice specifications that permit unobserved heterogeneity do so by assuming random coefficients. Examples are Hausman and Wise (1978), McFadden and Train (2000), Train (2009), and the demand side of Berry, Levinsohn, and Pakes (1995). Our model nests random coefficient models as a special case, with the added advantages that: 1. In our model the econometrician does not need to specify which regressors have random coefficients. 2. Our model permits the distribution of the random coefficients to depend on regressors. 3. The econometrician does not need to specify the functional form of the random coefficient distributions. 4. Our estimator remains numerically the same regardless of how many or few regressors have random coefficients, and 5. Our estimator does not require numerical integration or deconvolution techniques, which are often required in semiparametric random coefficient models. If random coefficients are present, our model assumes they are symmetrically distributed, and so for example nests the usual parametric assumption of normally distributed random coefficients as a special case.

Alternatively, in contrast to fixed coefficient parametric models of individual heterogeneity, such as Steckel and Vanhonacker (1988), Hensher, Louviere, and Swait (1999), or Fiebig, Keane,
Louviere, and Wasi (2009), our method allows the econometrician to be agnostic about the functional form of heteroskedasticity or about the number and types of component error structures if the latent variances in the population are given by a mixture of heterogeneous sub-populations.

Our identification and associated estimator does not require observing all the choices a consumer makes. Just observing whether each consumer chooses one particular choice or not suffices. So, e.g., we could identify all the parameters in a multinomial product choice model, even if we only observed whether each consumer bought some product or not, without seeing which product they chose. Observing and making use of other choices provides overidentifying information that would in practice be used to increase efficiency, or to test the model.

Like most semiparametric multinomial choice models, we require excluded regressors for point identification. That is, for each choice we assume there is a regressor that directly affects the utility of that choice and not of other choices, and is conditionally independent of the latent errors. However, we do not impose the typically required “large support” or special regressor assumption on these excluded regressors, as is assumed in, e.g., Manski (1975, 1985), Horowitz (1992), Matzkin (1993), Lewbel (2000), Berry and Haile (2010), and Fox and Gandhi (2016).  

The key identifying assumption we make is error symmetry, that is, we assume that the joint latent error distribution, conditional on covariates, is centrally symmetric. Though error symmetry has not previously been used for identification and estimation of multinomial discrete choice models, its identifying power has been investigated in binary choice models. Manski (1988) shows that conditional symmetry in binary response models does not have identifying power beyond median independence. Chen, Khan, and Tang (2016) in contrast find that symmetry, when combined with conditional independence of one regressor, does improve rates of convergence. Other studies, like Chen (2000) and Chen and Zhou (2010), use symmetry to improve efficiency.

---

1Kashaev (2018) also relaxes the full support condition on the excluded covariates in a discrete choice analysis, but requires parametric restrictions on the random coefficients for identification.

2A large class of multivariate distributions are centrally symmetric. See, e.g., Serfling (2006). A partial list includes multivariate normal distributions, multivariate logistic, multivariate Student’s t-distributions, multivariate Cauchy distributions, and mean zero mixtures of multivariate normal distributions. MNP models with or without normal random coefficients have conditionally centrally symmetric latent errors.
As we discuss in section 2, the method of using symmetry for identification in binary choice models does not immediately extend to the multinomial setting, because we must account for possible correlation in latent errors of different choices.

Symmetry has also been used to obtain point identification in a variety of other econometric models. Examples include the censored and truncated regression models of Powell (1986a, 1986b), the type 3 Tobit model of Honoré, Kyriazidou, and Udry (1997), stochastic frontier models as in Kumbhakar and Lovell (2000), omitted variable models as in Dong and Lewbel (2011), and measurement error models as in Lewbel (1997), Chen, Hu, and Lewbel (2008), and Delaigle and Hall (2016). These examples are all univariate dependent variable models. A previous example of employing joint symmetry to identify a multiple dependent variable model is the two player entry game of Zhou (2020).\(^3\)

An alternative method for introducing general heteroskedasticity into semiparametric discrete choice models is maximum score estimation. For multinomial (as opposed to binary) choice maximum score estimation, Manski (1975, 1985) requires independent, identically distributed errors. Fox (2007) relaxes this condition to exchangeable errors, but these methods still cannot handle the dependence across errors and the variability of random coefficients of the more general forms of heteroskedasticity that our method permits.\(^4\)

Our model is standard multinomial choice, where the utility of each choice is a linear index in covariates plus a possibly heteroskedastic latent error term that may be correlated with the latent errors of the other choices. Our goal is identification and estimation of the utility index coefficients, but we also recover the distribution of the latent errors, and hence choice probabilities.

Designating any one possible choice as the base or outside option, the minimum data required

---

\(^3\) We use a similar central symmetry condition as Zhou (2019), but otherwise differ in a variety of ways. Among other differences, our model is general multinomial choice, rather than a specific entry game model; we allow for an arbitrary number of choices, as opposed to just two players; we exploit symmetry of differences in utility rather than levels of payoffs to obtain identification; and we obtain a new and different root \(N\) consistent and asymptotically normal estimator based on our identification.

\(^4\) Yan and Yoo (2019) show that a generalized maximum score method can accommodate random coefficients, but they require fully rank-ordered choice data, rather than just the first choice as in standard multinomial discrete choice estimation.
to implement our identification strategy and associated estimator is the following. For each individual in the sample, we must observe the difference between the value of covariates for each choice vs the outside option, and an indicator of whether the individual chooses the outside option or not. It is not necessary to observe other choices of the individual, though doing so provides overidentifying information that can be used to increase efficiency or to test the model. Other semiparametric multinomial choice models that have this unusual feature of only requiring observation of a subset of the available choices are Lewbel (2000) and Fox (2007).

The intuition of our identification is as follows. In our setting, the expected value of making any one choice (which we may arbitrarily designate as the base option), conditional on covariates, equals the conditional distribution function of the latent errors, evaluated at the unknown values of the utility indices. Taking the derivatives of this function with respect to the excluded regressors yields the probability density function of the latent errors, also evaluated at the unknown utility index values. Conditional symmetry of the latent errors means that, at a given value of the covariates, we can construct a corresponding symmetry point that has the same value of the latent error conditional density function. Equating the estimated densities (which are just nonparametric regression derivatives) at these pairs of points provides equality restrictions on the utility indices that we use to identify the utility index parameters.

Using the analogy principle, we construct a corresponding estimator that minimizes the squared differences of the estimated error densities at each data point with its corresponding symmetry point. We show this minimum distance estimator is root N consistent and asymptotically normal. Computing the objective function of our estimator does not entail either numerical integration or deconvolution techniques, which are often required by random coefficients models. Moreover, our estimator does not require specifying which covariates, if any, have random coefficients, and is no more or less complicated regardless of how many covariates have random

\footnote{For example, in a purchase decision where regressors include prices of each choice, we must observe the difference between the price of each choice and the price of the outside option. If the outside option is making no purchase, then the price of the outside option would be zero. In this case we could identify the model even if we only observed whether each consumer made any purchase or not, without observing which product was purchased.}
coefficients, or any other more complicated forms of heteroskedasticity.

Many methods have been developed for identifying and estimating utility function parameters with cross-sectional multinomial choice data. Other examples, in addition to those cited above, include Ruud (1986), Lee (1995), Powell and Ruud (2008), Yan (2013), Fox and Gandhi (2016), Ahn, Ichimura, Powell, and Ruud (2018), Shi, Shum, and Song, (2018), and Khan, Ouyang, and Tamer (2019). Many of these estimators assume independence between the covariates and error terms, ruling out the possibility of individual heterogeneity, including Ruud (1986), Powell and Ruud (2008), Shi, Shum and Song (2018), and Khan, Ouyang, and Tamer (2019). Some assume exchangeable errors across alternatives, such as Manski (1975), Fox (2007), and Yan (2013). Exchangeability allows for flexible interpersonal heteroskedasticity, but greatly restricts the permitted forms of correlation and heteroskedasticity across alternatives. Lee (1995) and Ahn, Ichimura, Powell, and Ruud (2018) allow for a limited form of heteroskedasticity, where the conditional distribution of the error terms can depend on linear indices of covariates, instead of on the covariates themselves. All of these approaches in general exclude random coefficients. Lewbel (2000), Berry and Haile (2010), and Fox and Gandhi (2016) propose semiparametric methods that can accommodate random coefficients, but they require strong support restrictions on special regressors and on unobservables, and their estimators have slower than parametric rates of convergence.

The paper is organized as follows. We present the model and show identification in Section 2. In Section 3 we propose our minimum distance estimator, and derive its asymptotic properties. Monte Carlo simulations on the finite sample performance of our estimators are reported in Section 4, and Section 5 concludes. Proofs of our theorems are given in the Appendices.

---

2 The Model and Identification

We use the following notation. Random variables are written in lower-case italics, random vectors in lower-case bold italics, and random matrices in upper-case bold italics. All vectors are column vectors. \( \mathbb{R}^q \) denotes the \( q \)-dimensional Euclidean space, where \( q \) is a positive integer. \( \mathbf{0}_q \) (\( \mathbf{1}_q \)) denote a \( q \)-vector of zeros (ones). For any two vectors \( \mathbf{a} \equiv (a_1, \ldots, a_q)' \) and \( \mathbf{b} \equiv (b_1, \ldots, b_q)' \) in \( \mathbb{R}^q \), the inequality \( \mathbf{a} \leq \mathbf{b} \) holds if and only if \( a_k \leq b_k \) holds for every \( k = 1, \ldots, q \), where \( a_k \) and \( b_k \) are the \( k^{th} \) elements of vectors \( \mathbf{a} \) and \( \mathbf{b} \), respectively. We reserve letters \( j \), \( k \) and \( l \) for indexing alternatives, and the letter \( n \) for indexing observations. Where the distinction needs emphasis, we use \( x_{nj} \) to denote the \( n^{th} \) observation of the random vector \( x_j \). Letters \( P \) and \( E \) denote a probability and an expectation, respectively. The symbol \( ' \) denotes matrix transposition, \( \setminus \) denotes set difference, and \( \perp \) denotes statistical independence. The function \( 1(\cdot) \) is the indicator function that equals one when the event in the brackets is true and zero otherwise.

2.1 The Random Utility Framework

We consider a standard random utility model. An individual in the population of interest faces a finite number of alternatives and must choose one of them to maximize her utility. Let \( \mathbb{J} \equiv \{0, 1, \ldots, J\} \) denote the set of alternatives, where integer \( J \geq 2 \). Let \( \bar{z}_j \in \mathcal{R} \) and \( \bar{x}_j \in \mathbb{R}^q \) denote covariates that affect the utility of alternative \( j \) (the tilde is used here because later we’ll use simpler notation, omitting the tilde, to denote differences of these covariates). The (latent) utility \( \bar{u}_j \) from choosing alternative \( j \in \mathbb{J} \) is assumed to be given by:

\[
\bar{u}_j = \bar{z}_j \gamma^o + \bar{x}_j' \theta^o + \bar{\varepsilon}_j \quad \forall \, j \in \mathbb{J},
\]

where \( \gamma^o \in \mathcal{R} \) and \( \theta^o \in \mathbb{R}^q \) are the preference parameters of interest, and \( \bar{\varepsilon}_j \in \mathcal{R} \) is the unobserved random component of utility for alternative \( j \). The utility index \( \bar{z}_j \gamma^o + \bar{x}_j' \theta^o \) is often called systematic (or deterministic) utility, as opposed to the error term, \( \bar{\varepsilon}_j \), which is the unsystematic
(or stochastic) component of utility.

For each alternative \( j \in \mathbb{J} \), let a dummy variable, \( y_j \), indicate whether alternative \( j \) yields the highest utility among all the alternatives, that is,

\[
y_j = 1(\tilde{u}_j \geq \tilde{u}_k \quad \forall \ k \in \mathbb{J} \setminus \{j\}).
\] (2)

The choice of the individual is denoted \( y \equiv (y_0, y_1, \ldots, y_J) \), where \( \sum_{j=0}^{J} y_j = 1 \). The econometrician observes the covariates \( \tilde{z}_j \) and \( \tilde{x}_j \) for \( j \in \mathbb{J} \) (or at least differences in these covariates as discussed later). The econometrician also observes the choice \( y_0 \), and might observe other elements of \( y \) as well. The latent utility vector \( \tilde{u} \equiv (\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_J) \) is not observed.\(^7\)

As is common in semiparametric discrete choice modeling, identification of the parameters \( \gamma^o \) and \( \Theta^o \) requires a scale normalization since these coefficient are only unique up to a scale.\(^8\) We set \( \gamma^o = 1 \) as our scale normalization and focus on the identification and estimation of the parameter vector \( \Theta^o \in \mathbb{R}^q \) in the remainder of the paper. This normalization is without loss of generality assuming \( \gamma^o \neq 0 \), which can be tested by examining the significance of \( \tilde{z}_j \) in a nonparametric regression of any \( y_j \) on all the covariates. Also, if the covariate \( \tilde{z}_j \) represents a cost rather than a benefit of choosing alternative \( j \), and so negatively affects utility (which one could learn by observing the sign of \( \tilde{z}_j \) in a nonparametric regression of any \( y_j \) on covariates), then we can replace the variable \( \tilde{z}_j \) with its negative value \( -\tilde{z}_j \) in (1) without changing the remaining analysis.

The random utility function (1) can accommodate both alternative-specific and individual-specific covariates. To see this point, consider a utility function that spells out the distinction in these two types of covariates explicitly

\[
\tilde{u}_j = \tilde{z}_j + \tilde{w}'_j \beta^o + \tilde{s}' \alpha^o_j + \tilde{\varepsilon}_j \quad \forall \ j \in \mathbb{J},
\] (3)

\(^7\)To achieve point identification, we later impose continuity conditions on the covariate \( \tilde{z}_j \) for each alternative \( j \), which makes utility ties occur with zero probability. For this reason, we ignore utility ties throughout the paper.

\(^8\)Scale normalization is needed because multiplying every \( \tilde{u}_j \) in (2) by any positive number will not change the observed choices of any individual. See Lewbel (2019) regarding normalizations in general and discrete choice scale normalizations in particular.
where $\tilde{z}_j \in \mathcal{R}$ and $\tilde{w}_j \in \mathcal{R}^{q_1}$ denote the covariates that vary across alternatives (e.g., product attributes), and let $\tilde{s} \in \mathcal{R}^{q_2}$ include both a constant term and covariates that vary across individuals but not across alternatives (e.g., socioeconomic variables). Since only differences in utilities matter for determining choices, adding a constant to each $\tilde{u}_j$ will not change the observed choice $y_j$ defined in (2). We may therefore without loss of generality impose $\alpha'_0 = 0_{q_2}$ as a location normalization.

To illustrate (3), in the classical application of choice of where to shop (McFadden, 1974), $\tilde{z}_j$ could be the “price” of a shopping trip (a function of travel time and cost), which varies across both consumers and by store choice, $\tilde{w}_j$ could include other covariates that vary across alternatives such as store quality, and $\tilde{s}$ could include socioeconomic variables which do not vary with store choice, like the consumer’s age and gender. Interaction terms of consumer attributes with store attributes would be included in $\tilde{w}_j$. For identification, we require at least one covariate, the one we denote by $\tilde{z}_j$, to vary by both consumer and by store choice. This is the excluded regressor, because it only appears in the utility for choice $j$ and is excluded from the other choices.

When an application includes individual specific variables $\tilde{s} \in \mathcal{R}^{q_2}$ as covariates, for notational compactness we can construct a parameter vector $\theta^o \in \mathcal{R}^q$ and covariate vector $\tilde{x}_j \in \mathcal{R}^q$ where $q = q_1 + J \times q_2$ such that $\tilde{x}_j^t \theta^o = \tilde{w}_j^t \beta^o + \tilde{s}^t \alpha^o_j$ for each $j \in \mathcal{J}$. To be more specific, let $\theta^o$ be a vector that is partitioned into $J + 1$ blocks, where the first block is $\beta^o \in \mathcal{R}^{q_1}$, and where block $(j + 1)$ is $\alpha^o_j \in \mathcal{R}^{q_2}$ for $j = 1, \ldots, J$. Let $\tilde{x}_0$ denote a vector that is partitioned into $J + 1$ blocks, where the first block is $\tilde{w}_0 \in \mathcal{R}^{q_1}$ and each of the remaining $J$ blocks is $0_{q_2}$. For $j \in \mathcal{J} \setminus \{0\}$, let $\tilde{x}_j$ denote a vector that is partitioned into $J + 1$ blocks, where the first block is $\tilde{w}_j$, the $(j + 1)^{th}$ block is $\tilde{s}$, and each of the remaining $J - 1$ blocks is $0_{q_2}$. For example, when $J = 2$, we define

$$
\theta^o = \begin{pmatrix}
\beta^o \\
\alpha'^o_1 \\
\alpha'^o_2
\end{pmatrix},
\tilde{x}_0 = \begin{pmatrix}
\tilde{w}_0 \\
0_{q_2}
\end{pmatrix},
\tilde{x}_1 = \begin{pmatrix}
\tilde{w}_1 \\
\tilde{s}
\end{pmatrix},
\tilde{x}_2 = \begin{pmatrix}
\tilde{w}_2 \\
0_{q_2} \\
\tilde{s}
\end{pmatrix}.
$$

9
With this notation, we can write equation (3) in the form of equation (1). If no individual specific variables \( s \) are included in the model, then \( s = 1 \) (a scalar) and each \( \alpha_j^o \) is an alternative-specific constant for \( j = 1, \ldots, J \).

Only differences in utilities matter in making choice decisions. Without loss of generality, we set alternative \( 0 \in J \) as the base alternative (i.e., as the so-called outside option) and subtract the utilities of other alternatives by that of the base alternative. We normalize this outside option to have utility \( \tilde{u}_0 = 0 \). Denote the (location-normalized) utility vector \( u \equiv (u_1, \ldots, u_J)' \in \mathcal{R}^J \), where \( u_j = \tilde{u}_j - \tilde{u}_0 \). For each alternative \( j = 1, \ldots, J \), we have the utility function

\[
 u_j = z_j + x_j'\theta^o + \varepsilon_j,
\]

where \( z_j \equiv \tilde{z}_j - \tilde{z}_0, \ x_j \equiv \tilde{x}_j - \tilde{x}_0 \in \mathcal{R}^q \), and \( \varepsilon_k \equiv \tilde{\varepsilon}_k - \tilde{\varepsilon}_0 \). The location-normalized utility vector can be expressed as

\[
 u = z + X\theta^o + \varepsilon,
\]

where \( z \equiv (z_1, \ldots, z_J)' \in \mathcal{R}^J, \ X \equiv (x_1, \ldots, x_J)' \in \mathcal{R}^{J \times q}, \) and \( \varepsilon \equiv (\varepsilon_1, \ldots, \varepsilon_J)' \in \mathcal{R}^J \). We use the compact expressions of utility functions defined in (4) and (5) throughout the paper.

In some contexts, like product choice where \( j = 0 \) corresponds to not purchasing any product, it is commonly assumed that \( \tilde{z}_0 \) and \( \tilde{x}_0 \) are zero, making \( z_j \) and \( x_j \) equal \( \tilde{z}_j \) and \( \tilde{x}_j \). Regardless, we only require that differences \( z \) and \( X \) be observed, and regularity conditions (e.g., continuity of \( z \)) are only be imposed on \( z_j \) and \( x_j \), not on \( \tilde{z}_j \) and \( \tilde{x}_j \). In addition to these covariates, our identification only requires that \( y_0 \) be observed, not the entire vector of outcomes \( y \). This is possible because \( z \) provides information about the other outcomes. Nevertheless, the associated estimators will be more efficient by observing and making use of more the elements of \( y \), since each additional outcome \( y_j \) one observes provides additional overidentifying information.
2.2 Key Conditions For Identification

Here we describe the identification of the parameter vector \( \theta^o \in \mathcal{R}^q \) in (5). As discussed in Section 2.1, \( \theta^o \) may include coefficients of alternative-specific covariates.

By equation (2), the probability of choosing the base alternative is

\[
P(y_0 = 1) = P(\bar{u}_1 - \bar{u}_0 \leq 0, \ldots, \bar{u}_J - \bar{u}_0 \leq 0) = P(\bar{u} \leq 0_J).
\]

(6)

It follows from equations (5) and (6) that the conditional probability of choosing the base alternative is

\[
P(y_0 = 1 | z, X) = P(\bar{\epsilon} \leq -z - X \theta^o | z, X) \equiv F_{\bar{\epsilon}} (-z - X \theta^o | z, X),
\]

(7)

where the right-hand side of (7) is the distribution function of the error vector \( \bar{\epsilon} \) evaluated at point \(-z - X \theta^o\), conditional on covariates \((z, X)\). Let sets \( S_z \subseteq \mathcal{R}^J \) and \( S_X \subseteq \mathcal{R}^{J \times q} \) denote the supports of the random vector \( z \) and random matrix \( X \), respectively. Let sets \( S_z(X) \) and \( S_{\bar{\epsilon}}(X) \) denote the supports of vectors \( z \) and \( \bar{\epsilon} \) conditional on the values of \( X \), respectively.

Next we state and discuss the two key assumptions on the covariates and error terms that we use for identification.

**Assumption 1:** Conditional on almost every \( X \in S_X \), the covariate vector \( z \) is independent of the error vector \( \bar{\epsilon} \), and the conditional distribution function of \( z \), \( F_z(\cdot | X) \), is absolutely continuous over its support \( S_z(X) \).

The conditional independence assumption between \( z \) and \( \bar{\epsilon} \) in Assumption 1 is also known as a distributional exclusion restriction (Powell, 1994, p. 2484). This assumption allows for interpersonal heteroskedasticity on a subset of covariates, i.e., \( X \), but not on all the covariates \((z, X)\). The classic troika of parametric models, i.e., the multinomial logit, the nested logit, and the multinomial probit models, all assume independence between all the covariates and the error terms, i.e., \((z, X) \perp \bar{\epsilon}\), which naturally implies the much weaker conditional independence
between \( z \) and \( \varepsilon \), \((z \perp \varepsilon) \mid X\), in Assumption 1. Many semiparametric multinomial choice models also impose this stronger independence assumption, e.g., Ruud (1986) and Powell and Ruud (2008). As discussed in the introduction, almost all semiparametric multinomial choice estimators impose at least Assumption 1.

Assumption 1 immediately implies \( P(\varepsilon \leq t \mid z, X) = P(\varepsilon \leq t \mid X) \) for any vector \( t \) in the support \( S_{\varepsilon}(X) \). An important implication of Assumption 1 is

\[
E(y_0 \mid z, X) = P(y_0 = 1 \mid z, X) = P(\varepsilon \leq -z - X\theta^o \mid X) \equiv F_{\varepsilon}(-z - X\theta^o \mid X),
\]

where the first equality in (8) holds because \( y_0 \) is binary, and the second equality holds by (7) and the conditional independence between \( z \) and \( \varepsilon \) in Assumption 1. The left side of (8) is identified, and can be readily estimated by nonparametric regression. This equation is useful because it relates an identified conditional expectation function to the unknown \( \theta^o \) and unknown distribution of \( \varepsilon \).

We will not do so, but suppose we had assumed that \( z \) had a large support, meaning it could take on any value in the support of \(-X\theta^o - \varepsilon\). Then \( z \) would be a vector of what Lewbel (2000) calls “special regressors.” Lewbel (2000) uses equation (8), large support, and an assumption that \( \varepsilon \) and \( X \) are uncorrelated to point identify \( \theta^o \). Other authors that use special regressors to identify multinomial choice models include Berry and Haile (2010), Fox (2015), and Fox and Gandhi (2016). The disadvantages of special regressor based identification are that the large support assumption is a strong restriction, and that multinomial choice estimators based on special regressors typically depend on identification at infinity, and so often have slower than parametric convergence rates.

Given these drawbacks, we do not impose a large support assumption, and we do not rely on identification at infinity arguments. We instead make the following symmetry assumption.

**Assumption 2:** For almost every \( X \in S_X \), the conditional distribution function \( F_{\varepsilon}(t \mid X) \) admits an absolutely continuous density function, \( f_{\varepsilon}(t \mid X) \), which is centrally symmetric about
the origin, i.e.,

\[ f_\varepsilon(t \mid X) = f_\varepsilon(-t \mid X), \quad (9) \]

for any vector \( t \in S_\varepsilon(X) \) where \( S_\varepsilon(X) \subseteq \mathcal{R}^J \).

Assumption 2 is an error symmetry restriction. Without loss of generality we assume that the point of symmetry is the origin, because any nonzero term could be absorbed into the intercept of the utility index as discussed following equation (3).

To see how we obtain identification based on Assumptions 1 and 2, consider taking the derivatives of both sides of (8) with respect to each elements of \( z \) and, evaluate the resulting function at the point \( (z = z^*, X = X^*) \) and the point \( (z = -z^* - 2X^*\theta, X = X^*) \), for some chosen values of \( z^*, X^* \), and \( \theta \). Note these values must be chosen so that \( X^* \in S_\varepsilon, z^* \in S_\varepsilon(X^*) \), and \( -z^* - 2X^*\theta \in S_\varepsilon(X^*) \). By Assumption 1 this yields the equations

\[
\frac{\partial^J E(y_0 \mid z = z^*, X = X^*)}{\partial z_1 \ldots \partial z_J} = f_\varepsilon(-z^* - X^*\theta^o \mid X = X^*) \times (-1)^J, \quad (10)
\]

and

\[
\frac{\partial^J E(y_0 \mid z = -z^* - 2X^*\theta, X = X^*)}{\partial z_1 \ldots \partial z_J} = f_\varepsilon(z^* + 2X^*\theta - X^*\theta^o \mid X = X^*) \times (-1)^J. \quad (11)
\]

Observe that the left sides of equations (10) and (11) are both identified, and can be readily estimated as nonparametric regression derivatives, given \( \theta \). It then follows from Assumption 2 that if \( \theta = \theta^o \), then the right sides of equations (10) and (11) are equal to each other. Therefore, define function \( d_0(\theta; z^*, X^*) \) as the difference between the left sides of equations (10) and (11),

\[ d_0(\theta; z^*, X^*) \equiv \frac{\partial^J E(y_0 \mid z = z^*, X = X^*)}{\partial z_1 \ldots \partial z_J} - \frac{\partial^J E(y_0 \mid z = -z^* - 2X^*\theta, X = X^*)}{\partial z_1 \ldots \partial z_J}. \quad (12) \]

Based on Assumptions 1 and 2, we have that if \( \theta = \theta^o \), then \( d_0(\theta; z^*, X^*) = 0 \). Given some regularity conditions, setting the function \( d_0 \) equal to zero at a collection of values of \( z^* \) and \( X^* \) provides enough equations to point identify \( \theta^o \).
2.3 An Alternative Identification Strategy

Existing binary choice estimators that make use of latent error symmetry, such as Chen (2000) and Chen, Khan, and Tang (2016), exploit the restrictions symmetry imposes on the distribution function of the latent errors, rather than on their density function as in the previous subsection. Equivalently, those binary choice estimators make use of the restrictions symmetry imposes on the condition expectation of $y_0$, rather than on derivatives of that expectation.

In this subsection we obtain restrictions that symmetry imposes on the distribution function of the latent errors in the multinomial choice setting, yielding an alternative identification strategy that uses condition expectations of $y_0$ instead of their derivatives. As discussed below, it turns out that (unlike binary choice) estimation based on this alternative identification strategy has some drawbacks in the multinomial setting, which is why we prefer our density based approach.

For simplicity, consider a three-choice setting, so $J = 2$. Let $[a, b]$ be a rectangle in the support of error vector. Point $a = (a_1, a_2)$ is the lower left vertex of this rectangle and $b = (b_1, b_2)$ is the upper right vertex. The probability of an error vector being in rectangle $[a, b]$ is the same as the probability of error vector being in rectangle $[-b, -a]$ under our central symmetry assumption. This is because each point $t$ in rectangle $[a, b]$ has a unique symmetric point $-t$ in rectangle $[-b, -a]$ and $f_\varepsilon(t \mid X) = f_\varepsilon(-t \mid X)$. We can easily verify that

$$
\int_{[a,b]} f_\varepsilon(t \mid X) \, dt = \int_{[a,b]} f_\varepsilon(-t \mid X) \, dt = \int_{[-b,-a]} f_\varepsilon(t \mid X) \, dt,
$$

where the first equality in (13) holds by Assumption 2 and the second one holds by changing of variables.\(^9\)

The integrals on both sides of equation (13) can be computed using the conditional distribution

\(^9\)Equation (13) also holds when $J > 2$, taking $[a, b]$ and $[-b, -a]$ to be centrally symmetric hyper-rectangles, and it holds for the binary choice model $J = 1$, where $[a, b]$ and $[-b, -a]$ reduce to symmetric intervals about the origin.
functions of the error vector. For example, consider the left-hand side integral:

\[ \int_{[a,b]} f_\varepsilon(t \mid X) \, dt = P(a \leq \varepsilon \leq b \mid X) \]

\[ = P(a_1 \leq \varepsilon_1 \leq b_1, a_2 \leq \varepsilon_2 \leq b_2 \mid X) \]

\[ = P(\varepsilon_1 \leq b_1, \varepsilon_2 \leq b_2 \mid X) - P(\varepsilon_1 < a_1, \varepsilon_2 \leq b_2 \mid X) \]

\[ + P(\varepsilon_1 \leq b_1, \varepsilon_2 < a_2 \mid X) - P(\varepsilon_1 < a_1, \varepsilon_2 < a_2 \mid X). \]  \hspace{1cm} (14)

Next consider a pair of individuals \( m \) and \( n \) with covariates \((z_m, X_m)\) and \((z_n, X_n)\), respectively, and assume \( z_m > z_n \) without loss of generality. Following the derivation in (14), we can express the probability

\[ P(-z_m - X_m \theta^\circ \leq \varepsilon \leq -z_n - X_n \theta^\circ \mid X_m = X_n) \]  \hspace{1cm} (15)

as the linear combination of four distribution functions. For each vector \( t \) in the support of the error vector, we can identify the error distribution \( P(\varepsilon \leq t \mid X = X_m) \) using estimatable conditional mean functions, because by (8)

\[ E(y_0 \mid z = -t - X_m \theta^\circ, X = X_m) = P(\varepsilon \leq t \mid X = X_m). \]  \hspace{1cm} (16)

Therefore, the probability of the error vector being in rectangle \([-z_m - X_m \theta^\circ, -z_n - X_n \theta^\circ] \) is identified. Similar to (15), we can also compute the probability

\[ P(z_n + 2X_n \theta - X_n \theta^\circ \leq \varepsilon \leq z_m + 2X_m \theta - X_m \theta^\circ \mid X_m = X_n) \]  \hspace{1cm} (17)

using conditional estimatable conditional means. When \( \theta = \theta^\circ \), rectangles \([-z_m - X_m \theta^\circ, -z_n - X_n \theta^\circ] \) and \([z_n + 2X_n \theta - X_n \theta^\circ, z_m + 2X_m \theta - X_m \theta^\circ] \) are centrally symmetric to each other so probabilities (15) and (17) are equivalent. This equality between these two probabilities provides an alternative identification strategy to our method in the previous section (12) that utilizes density functions.

Our density based identification and estimation entails matching points (that is, using \( f_\varepsilon(t \mid X) = f_\varepsilon(-t \mid X) \) at data points \( t \)) rather than matching rectangles as above. Matching points rather
than rectangles is also possible using distributions in the binary choice setting, but not for multinomial choice\textsuperscript{10}. In contrast, matching densities rather than distributions at points works for identifying and estimating both binary and multinomial choice.

We prefer to identify and estimate $\theta$ by matching each point in the data using densities, rather than by matching rectangles using distributions, for many reasons. First, equating error distribution rectangles involves more tuning parameters, since rectangles need to be chosen. Second, matching densities only requires finding enough points ($z = z^*, X = X^*$) in the data that have matches ($z = -z^* - 2X^*\theta, X = X^*$) that lie in the support of the covariates. In contrast, each matching rectangle requires finding an entire range of covariates that lie in the support and has a range of matches that also lie entirely in the support. Third, to gain efficiency we will later create more moments by replacing $y_0$ with different choices $y_j$. When matching density points, the same covariate values (points) that work for any one choice $j$ will also work for any other choice. The same is not true for matching distribution rectangles, because for rectangles each match entails pairs of observations rather than individual observations. Finally, the computation cost of estimation is lower for equating error densities than for distribution rectangles. For a sample of size $N$, we compute error densities at $2N$ points, while in contrast, using rectangles would entail computing the error distribution at $N(N-1)2^J$ points.

\textbf{2.4 Regularity Conditions For Identification}

Here we provide one set of conditions that, when combined with Assumptions 1 and 2, suffice to point identify $\theta^0$.

\textbf{Assumption 3:} The true parameter vector $\theta^0$ is in the parameter space $\Theta$, where $\Theta$ is a compact set in $\mathcal{R}^q$.

A compact parameter space is a standard assumption for a wide class of nonlinear models,\textsuperscript{10} For binary choice we have $J = 1$, making $\varepsilon$ and $t$ being scalars, and by symmetry we get $F_{\varepsilon}(t \mid X) = 1 - F_{\varepsilon}(-t \mid X)$ at data points $t$, but this equality does not hold for the multinomial case $J > 1$, when $\varepsilon$ and $t$ are vectors rather than scalars.
including semiparametric multinomial discrete choice models. See, e.g., Manski (1975), Lee (1995), Lewbel (2000), or Fox (2007). Assumption 3 implies that $\Theta$ is bounded, which is essential for our identification because we do not impose full support conditions on covariates $z$. In particular, if $\theta$ were not bounded then $X^*\theta$ could be unbounded, pushing $-z^* - 2X^*\theta$ outside the support of $z$, meaning that we’d have no points $z^*, X^*$ at which we could evaluate the function $d_0(\theta; z^*, X^*)$.

For notational simplicity, let $int(S_{(z,X)})$ denote the interior of the support $S_{(z,X)}$. Since some or all of the covariates in the matrix $X$ may be discrete random variables, let $X_c$ and $X_d$, respectively, denote the continuous and discrete covariates in $X$. Define

$$int(S_{(z,X)}) \equiv \{(z^*, X^*_c, X^*_d) \in S_{(z,X_c,X_d)} | (z^*, X^*_c) \in int(S_{(z,X_c)}(X^*_d)), X^*_d \in S_{X_d}\}. \quad (18)$$

Next we formally define the point identification condition for $\theta^o \in \Theta$, based on the $d_0$ function.

**Definition 2.1** For every vector $\theta \in \Theta$, define a set

$$D_0(\theta) \equiv \{(z^*, X^*) \in int(S_{(z,X)}) | (-z^* - 2X^*\theta, X^*) \in int(S_{(z,X)}), d_0(\theta; z^*, X^*) \neq 0\}. \quad (19)$$

The true parameter vector $\theta^o \in \Theta$ is point identified if $P([z, X] \in D_0(\theta)) = 0$ if and only if $\theta = \theta^o$, where $\theta \in \Theta$.

At the true parameter vector $\theta^o$, $D_0(\theta^o)$ is an empty set because $d_0(\theta^o; z^*, X^*) = 0$ for every $(z^*, X^*), (-z^* - 2X^*\theta^o, X^*) \in S_{(z,X)}$. To achieve identification, we require that the set $D_0(\theta)$ have positive probability measure for any $\theta$ in the parameter space other than $\theta^o$. The following assumptions give one set of conditions that suffice. In particular, Assumption 4 provides a subset of the support of covariates with positive measure on which the function $d_0(\theta; z^*, X^*)$ can be identified, while Assumption 5 ensures that symmetry points are unique.

**Assumption 4:** The following conditions on the covariates hold.

(a) For any constant vector $c = (c_1, \ldots, c_q)' \in \mathbb{R}^q$, $P(Xc = 0_I) = 1$ if and only if $c = 0_q$. 

17
(b) The joint density function of the continuous random variables in $\mathbf{X}$ is absolutely continuous and positive over its support. For every $\mathbf{X}^* \in S_{\mathbf{X}}$, the conditional density function of $\mathbf{z}$ given $\mathbf{X}$, $f_{\mathbf{z}}(\cdot \mid \mathbf{X} = \mathbf{X}^*)$, is absolutely continuous and positive over its support $S_\mathbf{z}(\mathbf{X}^*)$.

(c) For every $\mathbf{X}^* \in S_{\mathbf{X}}$, there exists a subset $\tilde{S}_\mathbf{z}(\mathbf{X}^*)$, where $\tilde{S}_\mathbf{z}(\mathbf{X}^*) \subseteq \text{int}(S_\mathbf{z}(\mathbf{X}^*))$, with positive measure such that $-\mathbf{z}^* - 2\mathbf{X}^* \mathbf{\theta} \in \text{int}(S_\mathbf{z}(\mathbf{X}^*))$ for every $\mathbf{z}^* \in \tilde{S}_\mathbf{z}(\mathbf{X}^*)$ and $\mathbf{\theta} \in \Theta$.

Assumption 4(a) is a standard full rank condition for identification, ruling out perfect collinearity among the regressors. If it is violated, then for some nonzero vector $\mathbf{c} \in \mathcal{R}^J$, $\mathbf{X} \mathbf{c} = \mathbf{0}_q$ for almost every $\mathbf{X}$, which implies that $\mathbf{X} \mathbf{\theta}^o = \mathbf{X} (\mathbf{\theta}^o + \lambda \mathbf{c})$ for almost every $\mathbf{X}$ and any constant $\lambda \in \mathcal{R}$. Consequently, we would fail to distinguish $\mathbf{\theta}^o$ from any vector $\mathbf{\theta}^o + \lambda \mathbf{c}$ that is also in the parameter space $\Theta$. Assumption 4(b) requires all the continuous covariates to have a positive joint density, so the conditional means in (8) can be identified at interior points in the support of the covariates. Assumption 4(c) guarantees that with a positive probability the values of the conditioning covariates on the left-hand side of (11) are in the interior of the support of the covariates, so that the function $d_0(\mathbf{\theta}; \mathbf{z}^*, \mathbf{X}^*)$ can be identified given any $\mathbf{\theta} \in \Theta$ for a positive measure of covariates.

**Assumption 5:** The following conditions on the error terms hold.

(a) For every $\mathbf{X}^* \in S_{\mathbf{X}}$, $\tilde{S}_\mathbf{e}(\mathbf{X}^*) \equiv \{-\mathbf{z}^* - \mathbf{X}^* \mathbf{\theta}^o \mid \mathbf{z}^* \in S_\mathbf{z}(\mathbf{X}^*)\} \cup \{\mathbf{z}^* + 2\mathbf{X}^* \mathbf{\theta} - \mathbf{X}^* \mathbf{\theta}^o \mid \mathbf{z}^* \in S_\mathbf{z}(\mathbf{X}^*), \mathbf{\theta} \in \Theta\}$ is a subset of the interior of the support $S_\mathbf{e}(\mathbf{X}^*)$.

(b) For every $\mathbf{X}^* \in S_{\mathbf{X}}$ and any constant vector $\mathbf{r} \in \mathcal{R}^J$, $f_\mathbf{e}(\mathbf{t} \mid \mathbf{X} = \mathbf{X}^*) = f_\mathbf{e}(\mathbf{r} - \mathbf{t} \mid \mathbf{X} = \mathbf{X}^*)$ for every $\mathbf{t} \in \tilde{S}_\mathbf{e}(\mathbf{X}^*)$ and $\mathbf{r} - \mathbf{t} \in \tilde{S}_\mathbf{e}(\mathbf{X}^*)$ if and only if $\mathbf{r} = \mathbf{0}_J$.

Assumption 5(a) ensures that the error density function on the right-hand sides of (10) and (11) are always evaluated at interior points of the error support $S_\mathbf{e}(\mathbf{X}^*)$ for every $\mathbf{X}^* \in S_{\mathbf{X}}$. Assumption 5(b) requires that the error density function has a unique (local) symmetry point.
over a subset of its support, $\tilde{S}_e(X^*)$. This assumption does not rule out densities having any flat sections at all, but it does limit the range of any such flat sections. So, e.g., a uniform distribution would be ruled out. Assumption 5 is what primarily guarantees that the function $d_0(\theta; z^*, X^*)$ is non-zero with positive probability when $\theta \neq \theta^o$.

Given these assumptions we obtain identification as follows. The proof of Theorem 2.1 is provided in Appendix A.

**Theorem 2.1** If Assumptions 1-5 hold, then the parameter vector $\theta^o \in \Theta$ is point identified by Definition 2.1.

### 2.5 Individual Heterogeneity and Random Coefficients

Our identifying assumptions do not refer specifically to random coefficients. In this section we provide sufficient conditions for our key assumptions to hold when unobserved heterogeneity takes the form of random coefficients.

For clarity, we now add a subscript $n$ where relevant, to designate a specific individual $n$. For estimation we will have a sample of observations $n = 1, \ldots, N$. The utility functions are then

$$u_{nj} = z_{nj} + x_{nj}' \theta_n + \epsilon_{nj} \text{ for } j = 1, \ldots, J, \quad (20)$$

where $\theta_n \in \mathcal{R}^q$ is the preference parameter vector for individual $n$. We have $u_{n0}$ normalized to be zero. Now decompose the parameter vector $\theta_n$ as $\theta_n = \theta^o + \delta_n$, where $\theta^o$ is the vector of the median of each random coefficient and $\delta_n$ is the vector of deviations of each parameter from its median. Our symmetry assumption implies that $\theta^o$ will also be the mean coefficients, as long as these means exist, but we don’t impose this existence.\(^{11}\)

\(^{11}\)Before scale normalization, we can have a random coefficient on $z_{nj}$, as long as the sign of the coefficient is strictly positive or negative, and the coefficient does not vary by $j$. If negative, replace $z_{nj}$ with $-z_{nj}$. Our previously discussed scale normalization is then equivalent to redefining $\theta_n$ and $\epsilon_{nj}$ by dividing these random coefficients and errors for individual $n$ by the random coefficient of $z_{nj}$. See, e.g., Lewbel (2019) for details on this same normalization in the context of special regressors.
We can rewrite the utility function (20) in the form of equation (4), as

\[ u_{nj} = z_{nj} + x'_{nj} \theta^o + \varepsilon_{nj} \quad \text{for} \ j = 1, \ldots, J, \]

where \( \varepsilon_{nj} = (\varepsilon_{nj} + x'_{nj} \delta_n) \), which is often called the composite error term in the presence of random coefficients. Let \((u_n, z_n, X_n, \varepsilon_n)\) denote individual \(n\)'s \((u, z, \varepsilon)\) and let \((\varepsilon_n, \delta_n)\) denote individual \(n\)'s \((\varepsilon, \delta)\), where \( \varepsilon \equiv (\varepsilon_1, \ldots, \varepsilon_J)' \) and \( \delta \equiv (\delta_1, \ldots, \delta_q)' \). We can express individual \(n\)'s utility vector in equation (5) as

\[ u_n = z_n + X_n \theta^o + \varepsilon_n, \]

where \( \varepsilon_n = \varepsilon_n + X_n \delta_n \). By Theorem 2.1, if this composite error vector \( \varepsilon_n \) satisfies Assumption 1 (Exclusion Restriction) and Assumption 2 (Central Symmetry), then \( \theta^o \) is point identified under the regularity conditions given by Assumptions 3-5. We now give sufficient conditions for Assumptions 1 and 2 to hold with random coefficients.

**Assumption 1-RC:** Conditional on almost every \( X \in S_X \), the covariate vector \( z \) is independent of \((\varepsilon, \delta)\), and the conditional distribution function of \( z, F_z(\cdot \mid X) \), is absolutely continuous over its support \( S_z(X) \).

By the definition of the composite error vector, \( \varepsilon = \varepsilon + X \delta \), Assumption 1 follows immediately from Assumption 1-RC, because conditional independence between \( z \) and \((\varepsilon, \delta)\) implies conditional independence between \( z \) and \( \varepsilon \).

**Assumption 2-RC:** For almost every \( X \in S_X \), the conditional distribution function of \((\varepsilon, \delta)\), \( F_{(\varepsilon, \delta)}(t_e, t_c \mid X) \), where \( t_e \in \mathcal{R}^J \) and \( t_c \in \mathcal{R}^q \), admits an absolutely continuous density function, \( f_{(\varepsilon, \delta)}(t_e, t_c \mid X) \), which is centrally symmetric about the origin, i.e.,

\[ f_{(\varepsilon, \delta)}(t_e, t_c \mid X) = f_{(\varepsilon, \delta)}(-t_e, -t_c \mid X), \]

for any vector \((t_e, t_c) \in S_{(\varepsilon, \delta)}(X)\).
Assumption 2-RC nests as a special case the random coefficients MNP model in which $(\epsilon, \delta)$ are assumed to be jointly normal and independent of all covariates. To show that Assumption 2-RC is a sufficient condition for Assumption 2, we need to verify that the composite error vector, $\varepsilon = \epsilon + X\delta$, satisfies conditional central symmetry, or equivalently,

$$P(\varepsilon < t \mid X) = P(\varepsilon > -t \mid X),$$

for any vector $t \in \mathcal{R}^J$ in the support of the composite error vector $S_{\varepsilon}(X)$. To show this, we have

$$P(\varepsilon < t \mid X) \equiv P(\epsilon + X\delta < t \mid X)$$

$$= \int_{\{t_x, t_c\}|t_x + X t_c < t} \int_{(\epsilon, \delta)} (t_x, t_c \mid X) \, d(t_x, t_c)$$

$$= \int_{\{t_x^{cs}, t_c^{cs}\}|t_x^{cs} - X t_c^{cs} < t} \int_{(\epsilon, \delta)} (-t_x^{cs}, -t_c^{cs} \mid X) \, d(t_x^{cs}, t_c^{cs})$$

$$= \int_{\{t_x^{cs}, t_c^{cs}\}|t_x^{cs} + X t_c^{cs} > -t} \int_{(\epsilon, \delta)} (t_x^{cs}, t_c^{cs} \mid X) \, d(t_x^{cs}, t_c^{cs})$$

$$= P(\epsilon + X\delta > -t \mid X) \equiv P(\varepsilon > -t \mid X),$$

where the third equality in (23) holds by a change of variables (where $t_x^{cs} = -t_x$ and $t_c^{cs} = -t_c$) and the fourth equality hold by Assumption 2-RC. Thus we have verified that Assumption 2-RC is a sufficient condition for Assumption 2.

While both Assumptions 1-RC and 2-RC restrict the relationship between covariates and unobservables, they do allow the joint distribution of all the unobservables $(\epsilon, \delta)$ to vary with covariates $X$. In contrast, typical random coefficients models, like random coefficients MNP and MNL models (e.g., Hausman and Wise 1978) assume much stronger independence conditions, such as $(z, X) \perp (\epsilon, \delta)$, which rules out individual heterogeneity in the distribution of $(\epsilon, \delta)$. Even in flexible semiparametric random coefficients models like Fox and Gandhi (2016), the usual assumption is $(\epsilon, \delta) \perp (z, X)$, ruling out the possibility that the distribution of random preferences may vary across sub-populations.

In addition to allowing dependence of $(\epsilon, \delta)$ on $X$, our assumptions also do not require thin tails or unimodality, unlike, e.g., normal random coefficient MNP or MNL models. Each element
of $\theta^o$ is the median of the corresponding random coefficient $\theta_n$. By Assumptions 1-RC and 2-RC, each element is also the mean of the random coefficient if that mean exists, and is also the mode if the random coefficient is unimodal.

As discussed earlier, another advantage of our model over most random coefficients models is that estimators of $\theta^o$ usually require the econometrician to know exactly which covariates have random coefficients and which do not. In such models, allowing many or all coefficients to possibly be random generally results in severe numerical difficulties. On the other hand, misspecification of which coefficients are random and which are not will generally result in biased parameter estimates. Our model and associated estimator avoids these issues.

Typical empirical applications of random coefficient multinomial choice estimators also usually assume independence between $\epsilon$ and $\delta$ to reduce the number of parameters one must identify and estimate, particularly since covariances between latent errors are generally poorly identified. Our method does not require estimating these covariances, or numerical integration, so we do not impose this independence requirement. The computational requirements of our method are not affected either by the presence or absence of these covariances, or by the number of coefficients in $\theta_n$ that are random.

One restriction we do impose is that we require one covariate in each choice $j$, $z_j$, not have a random coefficient. Setting the coefficient of some covariate $z$ equal to one is often a natural, economically meaningful normalization. For example, utility of choices are typically modeled as benefits minus costs. Benefits may be subjective and so vary heterogeneously as in random coefficients, while costs are often objective and fixed. In these cases $z$ would be a cost measure. Examples are willingness to pay studies where the benefits equal the willingness to pay, and consumer choice applications where $z_j$ is the price of choice $j$. (See e.g., Bliemer and Rose 2013 for more discussion and examples).\[^{12}\]

\[^{12}\]Many other semiparametric random coefficients choice models have either the same restriction, such as Lewbel (2000) and Berry and Haile (2005), or a comparable restriction.
Nevertheless, we could also assume that, before normalizing, the variable \( z \) has a random coefficient, provided that the random coefficient is the same random variable for all choices, and is positive (this latter restriction is a special case of the hemisphere condition required by semiparametric binary choice random coefficient estimators. See, e.g., Gautier and Kitamura 2013). This restriction is needed because we can’t allow renormalizations that would change any individual’s relative ranking of utilities. Note that in this case, we require our symmetry condition to hold after renormalization, not before.

In addition to not requiring knowledge of which covariates have random coefficients, we can detect which covariates have random coefficients, and estimate their distribution. Suppose Assumptions 1-RC, 2-RC, 3, 4, and 5 hold. Then the composite error vector is \( \varepsilon_n = \varepsilon_n + X_n(\theta_n - \theta^0) \). Suppose in addition that the distribution of \((\varepsilon, \delta)\) does not depend on \( X \), so the only source of heteroskedasticity is random coefficients. Then the distribution of \( \varepsilon \) only depends on any given element of \( X \) if that element of \( X \) has a random coefficient. Once we have identified \( \theta^0 \) using Theorem 2.1, then by equation (8), under mild regularity conditions, we can identify \( F_\varepsilon(t \mid X) = E(y_0 \mid z = -t - X \theta^0, X) \) when the vector \(-t - X \theta^0\) is any interior point in the support \( S_\varepsilon(X) \). So the function \( F_\varepsilon(t \mid X) \), over this support, can be estimated by nonparametric regression. We could then test if the estimated \( F_\varepsilon(t \mid X) \) depends on any given element of \( X \). Under these assumptions, an element of \( X \) has a random coefficient if and only if \( F_\varepsilon(t \mid X) \) depends on that element of \( X \).

### 2.6 Identification Using Multiple Choices

In Section 2.2, we identified the parameter vector \( \theta^0 \) using only derivatives of the conditional mean of \( y_0 \). Here we illustrate that identification can be achieved using the conditional mean of \( y_j \) for any \( j \in \mathbb{J} \). Later we will increase efficiency of estimation by combining the identifying moments based on each of the observed choices \( y_j \).
We now introduce some additional notation. For each \( j \in \mathbb{J} \), define \( X^{(j)} \) as the matrix that consists of differenced covariate vectors \( \tilde{x}_k - \tilde{x}_j \) for all \( k \in \mathbb{J} \) and \( k \neq j \). For example, when \( 1 < j < J \), \( X^{(j)} \equiv (\tilde{x}_0 - \tilde{x}_j, \ldots, \tilde{x}_{j-1} - \tilde{x}_j, \tilde{x}_{j+1} - \tilde{x}_j, \ldots, \tilde{x}_J - \tilde{x}_j)' \in \mathcal{R}^{J \times q} \). By this notation, we have \( X^{(0)} \equiv (\tilde{x}_1 - \tilde{x}_0, \ldots, \tilde{x}_J - \tilde{x}_0) = X \). In the same fashion, define \( z^{(j)} \in \mathcal{R}^J \) as the vector of differenced covariates \( \tilde{z}_k - \tilde{z}_j \) for all \( k \neq j \) and \( \varepsilon^{(j)} \in \mathcal{R}^J \) as the vector of differenced error terms \( \tilde{\varepsilon}_k - \tilde{\varepsilon}_j \) for all \( k \neq j \), respectively. By this definition, we have \( z^{(0)} = z \) and \( \varepsilon^{(0)} = \varepsilon \). Define \( u^{(j)} \in \mathcal{R}^J \) as the vector of differenced utilities \( \tilde{u}_k - \tilde{u}_j \) for all \( k \neq j \). Differenced utility vectors are then given by

\[
u^{(j)} = z^{(j)} + X^{(j)} \theta^o + \varepsilon^{(j)}. \tag{24}
\]

For \( j = 0 \), equation (24) is the same as equation (5).

The conditional probability of choosing alternative \( j \in \mathbb{J} \) is

\[
P(y_j = 1 \mid z^{(j)}, X^{(j)}) = P(\tilde{u}_k - \tilde{u}_j \leq 0 \ \forall k \in \mathbb{J} \setminus \{j\} \mid z^{(j)}, X^{(j)})
= P(u^{(j)} \leq 0_j \mid z^{(j)}, X^{(j)})
= P(\varepsilon^{(j)} \leq -z^{(j)} - X^{(j)} \theta^o \mid z^{(j)}, X^{(j)})
= F_{\varepsilon^{(j)}}(-z^{(j)} - X^{(j)} \theta^o \mid z^{(j)}, X^{(j)}), \tag{25}
\]

where the right-hand side of (25) is the distribution function of the error vector \( \varepsilon^{(j)} \) evaluated at the point \(-z^{(j)} - X^{(j)} \theta^o\), conditional on covariates \((z^{(j)}, X^{(j)})\). Let sets \( S_{z^{(j)}} \subseteq \mathcal{R}^J \) and \( S_{X^{(j)}} \subseteq \mathcal{R}^{J \times q} \) denote the supports of the random vector \( z^{(j)} \) and random matrix \( X^{(j)} \), respectively. Let sets \( S_{z^{(j)}}(X^{(j)}) \) and \( S_{\varepsilon^{(j)}}(X^{(j)}) \) denote the supports of vectors \( z^{(j)} \) and \( \varepsilon^{(j)} \) conditional on the values of \( X^{(j)} \), respectively.

**Proposition 2.1** If Assumption 1 holds, then for every \( j \in \mathbb{J} \) and conditional on almost every \( X^{(j)} \in S_{X^{(j)}} \), covariate vector \( z^{(j)} \) is independent of the error vector \( \varepsilon^{(j)} \), i.e., \((z^{(j)} \perp \varepsilon^{(j)}) \mid X^{(j)}\). The distribution function \( F_{z^{(j)}}(\cdot \mid X^{(j)}) \), is absolutely continuous over its support \( S_{z^{(j)}}(X^{(j)}) \).
Proposition 2.1 is an immediate result of the fact that there is a one-to-one correspondence between \(X^{(j)}\) and \(X\), \(z^{(j)}\) and \(z\), and \(\varepsilon^{(j)}\) and \(\varepsilon\), respectively. Hence we have

\[
E(y_j \mid z^{(j)}, X^{(j)}) = P(y_j = 1 \mid z^{(j)}, X^{(j)}) = P(\varepsilon^{(j)} \leq -z^{(j)} - X^{(j)}\theta^o \mid X^{(j)}) = \int f_{\varepsilon^{(j)}}(\varepsilon^{(j)} \mid X^{(j)}) d\varepsilon^{(j)} = E\left(\int f_{\varepsilon^{(j)}}(\varepsilon^{(j)} \mid X^{(j)}) \mid X^{(j)}\right),
\]

where the first equality in (26) holds because \(y_j\) is a dummy variable, and the second one holds by (25) and Proposition 2.1. For every \(t \in S_\varepsilon(X)\), let \(t^{(j)}\) denote the vector such that \(\varepsilon^{(j)} = t^{(j)}\) when \(\varepsilon = t\). Given any \(t^{(j)} \in S_\varepsilon^{(j)}(X^{(j)})\) we can calculate the conditional mean of \(y_j\) on the left-hand side of (26) at \(z^{(j)} = -t^{(j)} - X^{(j)}\theta^o\) as

\[
E(y_j \mid z^{(j)} = -t^{(j)} - X^{(j)}\theta^o, X^{(j)}) = P(\varepsilon^{(j)} \leq t^{(j)} \mid X^{(j)}) = \int f_{\varepsilon^{(j)}}(t^{(j)} \mid X^{(j)}) d\varepsilon^{(j)}.
\]

**Proposition 2.2** If Assumption 2 holds, then for every \(j \in J\) and almost every \(X^{(j)} \in S_{X^{(j)}}\), the conditional distribution function \(F_{\varepsilon^{(j)}}(t^{(j)} \mid X^{(j)})\) admits an absolutely continuous density function, \(f_{\varepsilon^{(j)}}(t^{(j)} \mid X^{(j)})\), which is centrally symmetric about the origin, i.e.,

\[
f_{\varepsilon^{(j)}}(t^{(j)} \mid X^{(j)}) = f_{\varepsilon^{(j)}}(-t^{(j)} \mid X^{(j)}),
\]

for any vector \(t^{(j)} \in S_{\varepsilon^{(j)}}(X^{(j)})\) where \(S_{\varepsilon^{(j)}}(X^{(j)}) \subseteq \mathbb{R}^J\).

To show Proposition 2.2, observe that for any \(t \in S_\varepsilon(X)\), we have \(\varepsilon = t\) if and only if \(\varepsilon^{(j)} = t^{(j)}\) by the one-to-one correspondence between \(\varepsilon\) and \(\varepsilon^{(j)}\). Therefore,

\[
f_{\varepsilon^{(j)}}(t^{(j)} \mid X^{(j)}) = f_\varepsilon(t \mid X) = f_\varepsilon(-t \mid X) = f_{\varepsilon^{(j)}}(-t^{(j)} \mid X^{(j)}),
\]

where the second equality in (29) holds by Assumption 2.

Now the remaining derivations mimic that of Theorem 2.1. Taking the \(J^{th}\) order derivatives of both sides of (26) with respect to each element of \(z^{(j)}\) and evaluating them at \((z^{(j)} = z^{(j)*}, X^{(j)} = \)
\(X^{(j)*}\) and \((z^{(j)} = -z^{(j)*} - 2X^{(j)*}\theta, X^{(j)} = X^{(j)*})\), respectively, we obtain the equations

\[
\partial^J E(y_j \mid z^{(j)} = z^{(j)*}, X^{(j)} = X^{(j)*})/\partial z_1^{(j)} \ldots \partial z_J^{(j)}
\]

(30)

\[
= f_{\varepsilon^{(j)}}(-z^{(j)*} - X^{(j)*}\theta^o \mid X^{(j)} = X^{(j)*}) \times (-1)^J
\]

and

\[
\partial^J E(y_j \mid z^{(j)} = -z^{(j)*} - 2X^{(j)*}\theta, X^{(j)} = X^{(j)*})/\partial z_1^{(j)} \ldots \partial z_J^{(j)}
\]

(31)

\[
= f_{\varepsilon^{(j)}}(z^{(j)*} + 2X^{(j)*}\theta - X^{(j)*}\theta^o \mid X^{(j)} = X^{(j)*}) \times (-1)^J.
\]

By symmetry, if \(\theta = \theta^o\) then the two error densities on the right-hand sides of (30) and (31) are identical, which implies equality of their left-hand sides. So for any vector \(\theta \in \Theta\) and \((z^{(j)*}, X^{(j)*}), (-z^{(j)*} - 2X^{(j)*}\theta, X^{(j)*}) \in S_{(z^{(j)}, X^{(j)})}\), define \(d_j(\theta; z^{(j)*}, X^{(j)*})\) as the difference of the left-hand sides of (30) and (31), that is,

\[
d_j(\theta; z^{(j)*}, X^{(j)*}) \equiv \partial^J E(y_j \mid z^{(j)} = z^{(j)*}, X^{(j)} = X^{(j)*})/\partial z_1^{(j)} \ldots \partial z_J^{(j)}
\]

(32)

\[-\partial^J E(y_j \mid z^{(j)} = -z^{(j)*} - 2X^{(j)*}\theta, X^{(j)} = X^{(j)*})/\partial z_1^{(j)} \ldots \partial z_J^{(j)}.
\]

which always equals zero when \(\theta = \theta^o\) and may be non-zero when \(\theta \neq \theta^o\).

Then, analogous to equation (19) in Definition 2.1, define

\[
D_j(\theta) \equiv \left\{ (z^{(j)*}, X^{(j)*}) \in \text{int} \left( S_{(z^{(j)}, X^{(j)})} \right) \right\}
\]

(33)

\[
\left. \left( -z^{(j)*} - 2X^{(j)*}\theta, X^{(j)*} \right) \right\} \in \text{int} \left( S_{(z^{(j)}, X^{(j)})} \right), d_j(\theta; z^{(j)*}, X^{(j)*}) \neq 0 \right\}.
\]

Recall that there is a one-to-one correspondence, respectively, between \(X^{(j)}\) and \(X, z^{(j)}\) and \(z, \varepsilon^{(j)}\) and \(\varepsilon\). For every \((z^*, X^*) \in \text{int}(S_{(z, X)})\) such that \((-z^* - 2X^*\theta, X^*) \in \text{int}(S_{(z, X)})\), we immediately have \((z^{(j)*}, X^{(j)*}) \in \text{int}(S_{(z^{(j)}, X^{(j)})})\) and \((-z^{(j)*} - 2X^{(j)*}\theta, X^{(j)*}) \in \text{int}(S_{(z^{(j)}, X^{(j)})})\), as well as \(d_j(\theta; z^{(j)*}, X^{(j)*}) = 0\) if and only if \(d_0(\theta; z^*, X^*) = 0\). Therefore, we can also use the choice probability of any alternative in the choice set to achieve identification, by replacing the set \(D_0(\theta)\) with \(D_j(\theta)\) in Definition 2.1 and Theorem 2.1 for any \(j \in J\).
3 A Minimum Distance Estimator and its Asymptotic Properties

3.1 Population Objective Functions for Estimation

Based on the identification strategy described in Section 2, we develop a minimum distance estimator (hereafter, MD estimator) for $\theta_0 \in \Theta$ using the identifying restriction functions

$$d_j(\theta_0; z^{(j)*}, X^{(j)*}) = 0$$

for $j = 0, \ldots, J$, where $d_j$ is defined by equation (32).

For each $j$, the function $d_j(\theta; z^{(j)*}, X^{(j)*})$ is well defined if both points $(z^{(j)*}, X^{(j)*})$ and $(-z^{(j)*} - 2X^{(j)*}\theta, X^{(j)*})$ are in the interior of the support of covariates, $S_{(z^{(j)},X^{(j)})}$. For this reason, we only wish to evaluate the function $d_j(\theta; z^{(j)*}, X^{(j)*})$ at such points. This can be achieved by multiplying each function $d_j(\theta; z^{(j)*}, X^{(j)*})$ by a trimming function of the form

$$\tau_j(z^{(j)}, X^{(j)}; \bar{\theta}, \theta) \equiv \varsigma_j(z^{(j)}, X^{(j)}) \varsigma_j(-z^{(j)} - 2X^{(j)}\bar{\theta}, X^{(j)}) \varsigma_j(-z^{(j)} - 2X^{(j)}\theta, X^{(j)})$$

where $X^{(j)}\bar{\theta}$ ( $X^{(j)}\theta$) gives the upper (lower) bound value that the index $X^{(j)}\theta$ can take. A simple choice for the function $\varsigma_j(\cdot)$ is

$$\varsigma_j(z^{(j)}, X^{(j)}) \equiv 1\left(|z^{(j)}| \leq c_1^{(j)}\right) \times 1\left(|X^{(j)}| \leq C_2^{(j)}\right),$$

where the absolute value of a vector or matrix, $| \cdot |$, is defined as the corresponding vector or matrix of the absolute values of each element, $c_1^{(j)} \in \mathcal{R}^J$ is a vector of trimming constants for the covariate vector $z^{(j)}$, and $C_2^{(j)} \in \mathcal{R}^{J \times q}$ is a matrix of trimming constants for the covariate matrix $X^{(j)}$ such that $(c_1^{(j)}, C_2^{(j)})$ is in the interior of the support of covariates $S_{(z^{(j)},X^{(j)})}$.

Denote $S_{z^{(j)}; \bar{\theta}, \theta}^{Tr} \left(X^{(j)}, \bar{\theta}, \theta\right)$ as the largest set of values $z^{(j)}$ given $\bar{\theta}$, $\theta$, and $X^{(j)}$, such that
\[ S_{z(j)}^{T_r} \left( X^{(j)}, \bar{\theta}, \theta \right) \subset \text{int} \left( S_{z(j)} \left( X^{(j)} \right) \right) . \] Next, we describe the regularity conditions on the trimming function in Assumption 6.

**Assumption 6:** The trimming function \( \tau_j \left( z^{(j)}, X^{(j)}; \bar{\theta}, \theta \right) \) is strictly positive and bounded on \( S_{z(j)}^{T_r} \left( X^{(j)}, \bar{\theta}, \theta \right) \times \text{int} \left( S_{X(j)} \left( X^{(j)} \right) \right) \), and is equal to zero on its complementary set for \( j = 0, \ldots, J \).

This trimming function ensures that we only consider observations having symmetrically reflected points that are in \( S_{z(j)}^{T_r} \left( X^{(j)}, \bar{\theta}, \theta \right) \) and \( \text{int} \left( S_{X(j)} \left( X^{(j)} \right) \right) \), in order to guarantee that our choice probabilities are well defined, and hence that our resulting estimators are well-behaved. In addition, this trimming will let us avoid boundary bias issues when nonparametrically estimating the derivatives of choice probabilities. This trimming function also does not depend on the particular estimate of \( \theta \), which simplifies derivation of the resulting estimator’s asymptotic properties. For notational conciseness, we denote \( \tau_j \left( z^{(j)}, X^{(j)} \right) \equiv \tau_j \left( z^{(j)}, X^{(j)}; \bar{\theta}, \theta \right) \) and \( S_{z(j), X(j)}^{T_r} \equiv S_{z(j)}^{T_r} \left( X^{(j)}, \bar{\theta}, \theta \right) \times \text{int} \left( S_{X(j)} \left( X^{(j)} \right) \right) \). The population objective function of our proposed MD estimator is

\[
Q_j (\theta) \equiv \frac{1}{2} E \left[ \tau_j \left( z_{n}^{(j)}, X_{n}^{(j)} \right) d_j \left( \theta; z_{n}^{(j)}, X_{n}^{(j)} \right) \right]^2 \tag{36}
\]

The sample objective function we define later replaces the expectation in (36) with a sample average and replaces \( d_j \left( \theta; z_{n}^{(j)}, X_{n}^{(j)} \right) \) defined by (32) with an estimator of this function.

The next theorem shows that the true parameter vector \( \theta^o \) is the unique minimizer in the parameter space for the population objective function, showing identification based on the population objective function. We provide the proof of Theorem 3.1 in the Supplementary Appendix.

**Theorem 3.1** If Assumptions 1-6 hold, then (i) \( Q_j (\theta) \geq 0 \) for any \( \theta \in \Theta \) and (ii) \( Q_j (\theta) = 0 \) if and only if \( \theta = \theta^o \).
Next, we derive the sample objective function based on population objective function and the asymptotic properties of the MD estimator.

### 3.2 The MD Estimator

In this section, we define the sample objective function of our MD estimator by the analogy principle. We specify the sampling environment below.

**Assumption 7:** \( \{(y_n, z_n, X_n), \text{ for } n = 1, \ldots, N\} \) is a random sample drawn from the infinite population distribution.

To ease notation, we denote the conditional means

\[
E(y_j | z^{(j)} = z_n^{(j)}, X^{(j)} = X_n^{(j)}) \equiv \varphi_j(z_n^{(j)}, X_n^{(j)}) \equiv \varphi_{j,o}(z_n^{(j)}, X_n^{(j)}),
\]

\[
E(y_j | z^{(j)} = -z_n^{(j)} - 2X_n^{(j)}\theta, X^{(j)} = X_n^{(j)}) \equiv \varphi_j(-z_n^{(j)} - 2X_n^{(j)}\theta, X_n^{(j)}) \equiv \varphi_{j,cs}(z_n^{(j)}, X_n^{(j)}, \theta),
\]

and function

\[
d_j(\theta; z_n^{(j)}, X_n^{(j)}) \equiv \frac{\partial^d E(y_j | z^{(j)} = z_n^{(j)}, X^{(j)} = X_n^{(j)})}{\partial z_1^{(j)} \cdots \partial z_j^{(j)}} \frac{\partial^d E(y_j | z^{(j)} = -z_n^{(j)} - 2X_n^{(j)}\theta, X^{(j)} = X_n^{(j)})}{\partial z_1^{(j)} \cdots \partial z_j^{(j)}}
\]

\[
\equiv \varphi_{j,o}^{(j)}(z_n^{(j)}, X_n^{(j)}) - \varphi_{j,cs}^{(j)}(z_n^{(j)}, X_n^{(j)}, \theta),
\]

where \( \varphi_{j,o}^{(j)}(z_n^{(j)}, X_n^{(j)}) \equiv \partial^d \varphi_{j,o}(z_n^{(j)}, X_n^{(j)}) / \partial z_1^{(j)} \cdots \partial z_j^{(j)} \) and \( \varphi_{j,cs}^{(j)}(z_n^{(j)}, X_n^{(j)}, \theta) \) is defined in the similar way as \( \varphi_{j,o}^{(j)}(z_n^{(j)}, X_n^{(j)}) \). Now, consider a leave-one-out (LOO) Nadaraya-Watson (NW) estimator for \( \varphi_{j,o,-n}^{(j)} \) as

\[
\varphi_{j,o,-n}^{(j)}(\cdot, \cdot) = \frac{1}{N-1} \sum_{m=1, m \neq n}^N \frac{y_m K_{h_x}(z_m^{(j)} - \cdot) K_{h_x}(X_m^{(j)} - \cdot)}{y_m K_{h_x}(z_m^{(j)} - \cdot) K_{h_x}(X_m^{(j)} - \cdot)}.
\]
where

\[ K_{hz}(z_m^{(j)} - \cdot) = \prod_{l=1}^{J} h_{z_l}^{-1}k\left(h_{z_l}^{-1}(z_m^{(j)} - \cdot)\right), \]

\[ K_{hx}(X_m^{(j)} - \cdot) = \prod_{l=1}^{J} \prod_{r=1}^{q} h_{x_{lr}}^{-1}k\left(h_{x_{lr}}^{-1}(X_m^{(j)} - \cdot)\right). \]

The properties of the kernel function \( k \) and those of the bandwidth \( h_z \equiv (h_{z_1}, \ldots, h_{z_J})' \) and \( h_x \equiv (h_{x_1,1}, \ldots, h_{x_1,q}, \ldots, h_{x_J,1}, \ldots, h_{x_J,q})' \) are defined in Assumptions 9 and 10 below, respectively. We can now define the LOO NW estimator \( \hat{\varphi}_{j,cs,-n} \) for \( \varphi_{j,cs} \) in the same fashion. The partial derivatives are rather tedious. Here we adopt the simplest estimator: the first term of the analytical derivatives, to simplify our analysis, which is the unbiased estimator for the derivative of choice probability.\(^\text{13}\)

Now we can construct the estimator for function \( d_j(\theta; z_n^{(j)}, X_n^{(j)}) \) in (36) as

\[ \hat{d}_{j,-n}(\theta; z_n^{(j)}, X_n^{(j)}) \equiv \hat{\varphi}_{j,o,-n}^{(j)}(z_n^{(j)}, X_n^{(j)}) - \hat{\varphi}_{j,cs,-n}^{(j)}(z_n^{(j)}, X_n^{(j)}, \theta). \]  \tag{37} \]

By replacing the expectation in \( Q_j(\theta) \) with its sample mean and function \( d_j(\theta; z_n^{(j)}, X_n^{(j)}) \) with its LOO estimator \( \hat{d}_{j,-n}(\theta; z_n^{(j)}, X_n^{(j)}) \), we define the MD estimator

\[ \hat{\theta} \in \arg \min_{\theta \in \Theta} Q_{Nj}(\theta), \]

where

\[ Q_{Nj}(\theta) = \frac{1}{2N} \sum_{n=1}^{N} \sum_{\tau_j} \left[ r_j(z_n^{(j)}, X_n^{(j)}) \hat{d}_{j,-n}(\theta; z_n^{(j)}, X_n^{(j)}) \right]^2. \]

\(^\text{13}\)Note that the full analytical derivatives of the choice probability are

\[ \frac{\partial^l a(w) b(w)}{\partial w_1 \cdots \partial w_J} = \sum_{\gamma=0}^{l} \sum_{s=\gamma,w(s)} \frac{\partial^s a(w)}{\partial w_1(s) \cdots \partial w_J(s)} \frac{\partial^{l-s} b(w)}{\partial w_1(s) \cdots \partial w_{J-s}(s)}, \]

where \( w(s) \) is any \( s \)-element in \( w \) and \( w(s) \) is remaining \( J-s \) element in \( w \) but not in \( w(s) \). \( \sum_{s=\gamma,w(s)} \) represents the summation of all possible combination of derivatives associated with \( w(s) \). Then, we have the derivative

\[ \frac{\partial^l \hat{\varphi}_j(z^{(j)}, X^{(j)})}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} = \frac{\partial^l \hat{g}_j(z^{(j)}, X^{(j)})}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \]

\[ = \sum_{\gamma=0}^{l} \sum_{s=\gamma,z(s)} \frac{\partial^s \hat{g}_j(z^{(j)}, X^{(j)})}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \frac{\partial^{l-s} \hat{f}_j(z^{(j)}, X^{(j)})}{\partial z_1^{(j)} \cdots \partial z_{J-s}(s)}, \]

30
We denote the gradient of the objective function as $q_{Nj}(\theta) = \nabla_\theta Q_{Nj}(\theta)$ and the Hessian matrix of the objective function as $H_{Nj}(\theta) = \nabla_{\theta\theta} Q_{Nj}(\theta)$. The smoothness of the objective function suggests the first-order condition (FOC): $q_{Nj}(\hat{\theta}) = 0$. Applying the standard first-order Taylor expansion to $q_{Nj}(\hat{\theta})$ around the true parameter vector $\theta^o$ yields

$$q_{Nj}(\hat{\theta}) = q_{Nj}(\theta^o) + H_{Nj}(\hat{\theta})(\hat{\theta} - \theta^o),$$

where $\hat{\theta}$ is a vector between the MD estimator $\hat{\theta}$ and the true parameter vector $\theta^o$. Then the influence function will be given by

$$\hat{\theta} - \theta^o = -\left[H_{Nj}(\hat{\theta})\right]^{-1} q_{Nj}(\theta^o). \quad (38)$$

We will show that $H_{Nj}(\hat{\theta}) \rightarrow_p H_j(\theta^o)$, where

$$H_j(\theta^o) = E \left\{ \tau_j^2 \left( z^{(j)}_n, X^{(j)}_n \right) \nabla_\theta d_j \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \left[ \nabla_\theta d_j \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \right]' \right\}, \quad (39)$$

and $\sqrt{N}q_{Nj}(\theta^o) \rightarrow_d N(0_q, \Omega_j)$, where $\Omega_j$ is the probability limit of the variance-covariance matrix of $q_{Nj}(\theta^o)$. To obtain these properties, we assume the following regularity conditions.

### 3.3 Regularity Conditions

Below, we will introduce some regularity conditions we will use to show the asymptotic normality.

**Assumption 8:** The following smoothness conditions hold:

(a) The density function $f_j \left( z^{(j)}, X^{(j)} \right)$ is continuous in the components of $z^{(j)}$ for all $z^{(j)} \in S_z^{Tr} \left( X^{(j)}, \overline{\theta}, \theta \right)$ and $X^{(j)} \in \text{int} (S_X)$. In addition, $f_j \left( z^{(j)}, X^{(j)} \right)$ is bounded away from zero uniformly over its support.
(b) Functions \( f_j(z^{(j)}, X^{(j)}) \), \( g_j(z^{(j)}, X^{(j)}) \) and \( \varphi_j(z^{(j)}, X^{(j)}) \) are \( s \) \((s \geq J + 1)\) times continuously differentiable in the components of \( z^{(j)} \) for all \( z^{(j)} \in S_{z}^{Tr} (X^{(j)}, \bar{\theta}, \theta) \) and have bounded derivatives.

**Assumption 9:** The kernel function \( k \) is an \( l \)-th \((l \geq 1)\) order bias-reducing kernel that satisfies

(a) \( k(u) = k(-u) \) for any \( u \) in the support of function \( k \) and \( \int k(u) du = 1 \);

(b) \( \int |u|^i k(u) du < \infty \) for \( 0 \leq i \leq l \);

(c) \( \int u^i k(u) du = 0 \) if \( 0 < i < l \) and \( \int u^i k(u) du \neq 0 \) if \( i = l \);

(d) \( k(u) = 0 \) for all \( u \) in the boundary of the support of kernel;

(e) \( \sup_u |k^{(1)}(u)|^2 < \infty \), where \( k^{(1)}(u) \) is the first derivative of \( k(u) \).

**Assumption 10:** The bandwidth vector \( h_z \equiv (h_{z_1}, \cdots, h_{z_J})' = (h_N, \cdots, h_N)' \) is a \( J \times 1 \) vector and the bandwidth \( h_X \equiv (h_{x_{1,1}}, \cdots, h_{x_{1,q}}, \cdots, h_{x_{J,1}}, \cdots, h_{x_{J,q}})' = (h_N, \cdots, h_N, \cdots, h_N, \cdots, h_N)' \) is a \( Jq \times 1 \) vector. The scalar \( h_N \) satisfies (a) \( h_N \to 0 \) and \( Nh_N^{2J+J+Jq} \to \infty \) as \( N \to \infty \); and (b) \( \sqrt{Nh_N^{2s}} \to 0 \) and \( \sqrt{N} (\ln N) \left( Nh_N^{2(J+1)+J+Jq} \right)^{-1} \to 0 \) as \( N \to \infty \).

**Assumption 11:** The components of the random vectors \( \partial^I \varphi_{j,o} / \partial z_1^{(j)} \cdots \partial z_J^{(j)} \), \( \partial^I \varphi_{j,cs} / \partial z_1^{(j)} \cdots \partial z_J^{(j)} \) and the random matrix \( \left( \partial^I \xi_j / \partial z_1^{(j)} \cdots \partial z_J^{(j)} \right) [y^{(j)}, z^{(j)}]' \) have finite second moments. Also, 

\[
\partial^I \varphi_{j,o} / \partial z_1^{(j)} \cdots \partial z_J^{(j)} , \partial^I \varphi_{j,cs} / \partial z_1^{(j)} \cdots \partial z_J^{(j)} \text{ and } \partial^I \xi_j \varphi_{j,o} / \partial z_1^{(j)} \cdots \partial z_J^{(j)}, \partial^I \xi_j \varphi_{j,cs} / \partial z_1^{(j)} \cdots \partial z_J^{(j)}
\]

where \( \xi_j \left( z^{(j)}, X^{(j)}, \theta^o \right) = r_j^2 \left( z^{(j)}, X^{(j)} \right) \nabla \theta d_j \left( \theta^o; z^{(j)}, X^{(j)} \right) \), satisfy the following Lipschitz
conditions: for some \( m(z^{(j)}, \cdot) \)

\[
\begin{align*}
\frac{\partial^j \phi_{j,o}(z^{(j)}, t_r)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}} &< m(z^{(j)}, \cdot) \| t_r \| \\
\frac{\partial^j \phi_{j,v}(z^{(j)}, t_r)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}} &< m(z^{(j)}, \cdot) \| t_r \| \\
\frac{\partial^j \xi_j \phi_{j,o}(z^{(j)}, t_r)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}} &< m(z^{(j)}, \cdot) \| t_r \| \\
\frac{\partial^j \xi_j \phi_{j,v}(z^{(j)}, t_r)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}} &< m(z^{(j)}, \cdot) \| t_r \|
\end{align*}
\]

with \( E [(1 + |y_j| + \| z^{(j)} \|) m(z^{(j)}, \cdot)]^2 < \infty. \)

**Assumption 12:** The matrix \( H_j \) defined by (39) is nonsingular and positive definite.

Assumption 8 gives the smoothness condition of the density and the choice probability. Assumption 9 collects restrictions for the kernel function. Assumption 10 describes the conditions on the bandwidth to achieve \( \sqrt{N} \) asymptotics. Assumption 11 imposes standard bounded moment and dominance conditions. Assumption 12 requires the hessian matrix is strictly positive definite.

Given these regularity assumptions, we will show asymptotic properties of the estimator.

### 3.4 Asymptotic Properties

The next two theorems establish the asymptotic properties of the MD estimator with proofs given in the Supplementary Appendix.

**Theorem 3.2** If Assumptions 1-10(a) hold, then the MD estimator \( \hat{\theta} \) converges to the true parameter vector \( \theta^o \in \Theta \) in probability.

**Theorem 3.3** Let Assumptions 1–12 hold. Then

(a) (Asymptotic Linearity) The MD estimator \( \hat{\theta} \) is asymptotically linear with

\[
\sqrt{N} \left( \hat{\theta} - \theta^o \right) = -N^{-1/2} \sum_{n=1}^{N} H_j^{-1} t_{nj} + o_p(1),
\]

33
where \( t_{nj} \equiv (v_{nj,o} - v_{nj,cs}) \partial^J \left[ \tau_n^2 \left( z_n^{(j)}, X_n^{(j)} \right) \nabla_{\theta} d_j \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) \right] / \partial z_{j}^{(j)} \ldots \partial z_{j}^{(j)} \) with scalars \( v_{nj,o} \equiv y_{nj} - \varphi_{j,o} \left( z_n^{(j)}, X_n^{(j)} \right) \) and \( v_{nj,cs} \equiv y_{nj} - \varphi_{j,cs} \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \).

(b) (Asymptotic Normality) The MD estimator is asymptotically normal, i.e.,

\[
\sqrt{N} \left( \hat{\theta} - \theta^o \right) \rightarrow_d N \left( 0, H_{j}^{-1} \Omega_j H_{j}^{-1} \right)
\]

where matrix \( \Omega_j \equiv E \left( t_{nj} t'_{nj} \right) \) and \( H_{j} \) is defined by (39).

In a supplementary appendix, we show that the leading term of the numerator of this estimator in equation (39) can be written as a U-statistic of order 2, and using the Hoeffding decomposition, we can decompose this U-statistic into a mean term, linear terms and a quadratic term. One linear term contributes to the limiting distribution, while the others are asymptotically negligible, following Newey and McFadden (1994) and Sherman (1993, 1994). Like the U-statistics in Powell, Stock and Stoker (1989), Newey (1994), and Imbens and Ridder (2009), our U-statistic is an average over a plug-in nonparametric estimator. We thereby achieve the parametric rate, which is unusual for semiparametric multinomial choice estimators.

Once all of our coefficients are estimated, the distribution of the model’s latent errors, and hence choice probabilities, can be identified and estimated as described at the end of section 2.5. This includes recovering the distributions of random coefficients if heteroskedasticity takes that form, and determining which covariates have random coefficients.

Our simplest MD estimator only requires observing a single choice \( j \) (e.g., selecting the outside option or not) and minimizing \( Q_{Nj} (\theta) \), which is a sample average of the square of the trimmed \( d_j \) function. If we observe more choices, we can instead minimize the sum of the \( Q_{Nj} (\theta) \) functions, summed over all observed choices \( j \). For efficiency, one could also consider minimizing a weighted sum. Moreover, since the expected value of the squared trimmed \( d_j \) function for each \( j \) is zero at the true \( \theta \), it would be possible to construct a generalized method of moments (GMM) estimator.
that minimizes a quadratic in the sample average of the vector of squared trimmed \( d_j \) functions for observed choices \( j \). However, because the elements of \( d_j \) are estimated derivatives of conditional expectations, the corresponding GMM second moment matrix would converge to a zero matrix, and as a result, standard GMM asymptotic theory would not apply. We therefore leave the question of efficient combination of \( Q_{Nj}(\theta) \) over multiple \( j \) for future research.

We conclude this section by discussing possible testing of our central symmetry assumption. Under the null hypothesis of a central symmetry, the error density at any two symmetric points would be equal, while under the alternative there must exist symmetric points where the densities are not equal. Also under the null, our estimator is consistent. So a test could be constructed based on the difference in error density estimates at many symmetry points (other than those used for estimation), using our estimated parameters to construct symmetry points. More general specification tests could also be constructed, using the fact that our parameters are over identified when more than one choice \( j \) is observed.

4 Monte Carlo Experiments

In this section, we use Monte Carlo experiments to study the finite-sample properties of the minimum distance (MD) estimator defined in Section 3. We consider four data generating processes (DGPs). In each DGP, individual \( n \)'s utility from alternative \( j \), \( u_{nj} \), is specified as

\[
u_{nj} = z_{nj} \gamma^o + x_{nj} \theta_n + \varepsilon_{nj} \text{ for } n = 1, 2, ..., N \text{ and } j = 0, 1, 2.
\] (40)

Each DGP is used to simulate two sets of 2000 random samples of \( N \) individuals, where sample size \( N = 1000 \) in the first set and \( N = 2000 \) in the second set.

In all DGPs, the first coefficient \( \gamma^o = 1 \). In DGPs 1 and 2, the second coefficient is also a constant \( \theta_n = \theta^o \) where \( \theta^o = 0.2 \), while in DGPs 3 and 4 the second coefficient is a random variable \( \theta_n = \theta^o + \delta_n \) with the distribution of \( \delta_n \) specified in Table 1. In DGP 3, the random coefficient \( \theta_n \) is independent of all the covariates. By contrast, the conditional distribution of \( \theta_n \) in DPG 4
varies with individuals’ characteristics. We normalize the scale of the first coefficient and focus on the estimation of $\theta^o$, which is the median and mean of the second coefficient $\theta_n$ under all DGPs. Details of the distribution of each DGP are provided in Table 1.

<table>
<thead>
<tr>
<th>DGP</th>
<th>Distribution of $\theta_n$</th>
<th>Distribution of $\varepsilon_{nj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\theta_n = 0.2$</td>
<td>$\varepsilon_{nj} = \varepsilon_{nj}$</td>
</tr>
<tr>
<td>2</td>
<td>$\theta_n = 0.2$</td>
<td>$\varepsilon_{nj} = \frac{1}{2} e^{2x_{nj}\varepsilon_{nj}}$,</td>
</tr>
<tr>
<td>3</td>
<td>$\theta_n = 0.2 + \delta_n$</td>
<td>$\varepsilon_{nj} = \frac{1}{2} \varepsilon_{nj}$</td>
</tr>
<tr>
<td></td>
<td>where $\delta_n = \frac{1}{2} \theta_n$</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$\theta_n = 0.2 + \delta_n$</td>
<td>$\varepsilon_{nj} = \frac{1}{2} \varepsilon_{nj}$</td>
</tr>
<tr>
<td></td>
<td>where $\delta_n = (e^{x_{n1}} + e^{x_{n2}}) \times \vartheta_n$</td>
<td></td>
</tr>
</tbody>
</table>

Note: both $\vartheta_n$ and $\varepsilon_{nj}$ are standard normal random variables, and they are independent of each other and all the covariates, and i.i.d. across the subscripted dimension(s).

The researcher observes attributes $z_{nj}$ and $x_{nj}$ for $j = 0, 1, 2$ and $n = 1, 2, ..., N$. For the MD estimator, we consider both the case where the researcher only observes whether the outside option (i.e., alternative 0) is chosen, and so just minimizes $Q_{N0}(\theta)$, and the case where the researcher also observes which alternative is chosen by each decision maker, and so minimizes the sum of $Q_{Nj}(\theta)$ for $j = 0, 1, 2$. In all DGPs, each covariate $z_{nj}$ is a continuous uniform random variable over the interval $[-9, 9]$ and $x_{nj}$ is a binary variable that takes value of 2 or -2 with equal probability for $j = 1, 2$. The covariates of alternative 0 are $z_{n0} = 0$ and $x_{n0} = 0$. All the observed covariates are independent of each other and i.i.d. across the subscripted dimension(s).

We use a grid search to compute our MD estimator over a parameter space of $[-0.8, 0.8]$ with the bin width of 0.05. In the estimation of choice probabilities we apply a truncated normal density for the kernel function $k_h(\cdot)$ with bandwidth $h_j = sd(z_{nj})N(-1/24)$, where $j = 1, 2$. Our
bandwidth is derived by minimizing the mean squared errors (MSE) of the second order derivative of the choice probability shown in Lemma S.A.4 in Supplement Appendix.\footnote{The bias and variance of the second derivative of the choice probability are $O(h^s)$ with $s \geq J + 1$ and $O(Nh_N^{-2(J+1)+J+Jq})$, respectively. Then, Silverman’s Rule of Thumb suggests that the optimal bandwidth is of order $N^{-1/(2s+2(J+1)+J+Jq)}$. In our simulation ($J = 2, s = 2J + 2$ and $q = 1$), we choose $h_j = cN^{(-1/22)}$ with $c = sd(z_{nj})$. Cross-validation is an alternative way to select the bandwidth but it is more computation-intensive. We leave it to the future studies to verify how our estimator performs across different choices of bandwidth.}

We compare the MD estimator with a flexible multinomial probit (MNP) estimator that has no constraint on the variance-covariance parameters of the underlying multivariate normal densities. Since we normalize the first coefficient, our MNP specification requires estimating the second preference parameter and three parameters of the error vector variance-covariance matrix.

Under DGP 1, the MNP model is correctly specified and the MNP estimator is efficient, so comparisons with the MNP estimates show us the efficiency loss that comes from using our robust semiparametric estimator. Under DGP 2, the MNP model is misspecified because the error and covariate vectors are not independent; the conditional variance matrix of the error vector depends on covariates. By contrast, our MD estimator remains consistent because conditional error symmetry holds. Under DGP 3, the MNP model is misspecified but the random coefficients MNP is correctly specified. Under DGP 4, the distribution of the random coefficient $\theta_n$ varies across sub-populations. In our design, the covariate vector $(x_{n1}, x_{n2})$ takes four possible different values, each yielding a different distribution of $\theta_n$. Both the MNP and the random coefficient MNP models are misspecified in this case, while our MD estimator remains consistent.

Table 2 reports the bias and root mean square error (RMSE) of each estimator in our simulations. The first set of columns reports the MNP estimator, the second reports our MD estimator using only observations of the outside option choice $y_0$, while the third uses observations of all choices $y_0, y_1, y_2$ (MNP also uses observations of all choices).

Under DGP 1, the MD estimators have small finite sample bias, and RMSEs two to four times larger than that of the correctly specified MNP estimator, which indicates the extent of efficiency loss of our estimator relative to the efficient MNP. Under DGP 2, the bias of the now misspecified
Table 2: Monte Carlo Results of estimating \( \theta^o \) (True Parameter \( \theta^o = 0.2 \))

<table>
<thead>
<tr>
<th>DGP</th>
<th>N</th>
<th>Bias</th>
<th>RMSE</th>
<th>Bias</th>
<th>RMSE</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000</td>
<td>-0.0012</td>
<td>0.0435</td>
<td>0.0216</td>
<td>0.2368</td>
<td>-0.0017</td>
<td>0.1337</td>
</tr>
<tr>
<td>1</td>
<td>2000</td>
<td>-0.0010</td>
<td>0.0307</td>
<td>0.0055</td>
<td>0.1355</td>
<td>-0.0078</td>
<td>0.0788</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>0.5656</td>
<td>0.5833</td>
<td>0.1047</td>
<td>0.3521</td>
<td>-0.0392</td>
<td>0.3048</td>
</tr>
<tr>
<td>3</td>
<td>2000</td>
<td>0.5627</td>
<td>0.5714</td>
<td>0.0543</td>
<td>0.2308</td>
<td>-0.0289</td>
<td>0.1747</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td>-0.7512</td>
<td>0.7718</td>
<td>-0.0054</td>
<td>0.3765</td>
<td>-0.0748</td>
<td>0.3550</td>
</tr>
<tr>
<td>5</td>
<td>2000</td>
<td>-0.7481</td>
<td>0.7585</td>
<td>0.0180</td>
<td>0.2616</td>
<td>-0.0343</td>
<td>0.2149</td>
</tr>
</tbody>
</table>

MNP estimator is around three times the true parameter value, and this bias remains as the sample size is doubled. In contrast, the bias and RMSE of the MD estimators are much smaller than the MNP estimator, and they decrease sharply as the sample size increases. In DGP 3, the random coefficients MNP is correctly specified, and so performs better than the MD estimators in terms of bias and RMSE. However, when the random component is heterogeneous in DGP4, the bias of MNP is almost four times the true parameter value and does not vanish as sample size grows. In contrast, the bias of the MD estimators is still a relatively small magnitude.\(^{15}\) In all the DGPs, in terms of RMSE, the MD estimator using all the choice information performs better than the MD estimator that only relies on the information regarding the outside option choice.

Our Monte Carlo experiments study the finite sample performance of our semiparametric method, but the results also provide evidence regarding the reliability of the multinomial probit model, which is generally considered to be a very robust parametric estimator, since it relaxes the restrictive error structure of the popular multinomial logit and nested logit estimators (Hausman and Wise, 1978; Goolsbee and Petrin, 2004). In multinomial discrete choice, both unobserved choice attributes and individual heterogeneity add complexity to the error structure. Our results

\(^{15}\) We speculate that the bias in the MD estimators might be further reduced by a bandwidth search, and/or using local linear estimation for the first stage choice probabilities.
show that ignoring either one may result in severely biased estimation.

5 Conclusion

We propose a new semiparametric identification and estimation method for the multinomial discrete choice model, based on error symmetry. This allows for very general heteroskedasticity across both individuals and alternatives, and general covariate dependent error correlations among alternatives. We do not assume the existence of error moments, or independence between covariates and errors, nor do we require large support assumptions or identification at infinity arguments.

As a result, heavy-tailed errors and random coefficients are permitted. Utilizing error symmetry, we propose an M-estimator that minimizes the squared difference of the estimated error density over pairs of symmetric points. We show that the estimator is root-N consistent and asymptotically normal. Monte Carlo experiments demonstrate finite-sample performance of the estimator under various DGPs, and compares favorably to multinomial probit models.

Our study opens a few promising areas to explore. Our model can readily incorporate control function type endogeneity, in the usual way of including estimated control function residuals as additional regressors, as in Blundell and Powell (2004). An open question is whether our method can be extended to allow for simultaneously determined prices as in the so-called micro-BLP model of Berry, Levinsohn, and Pakes (2004) or Berry and Haile (2010).

A Appendix: Proof of Identification

Proof of Theorem 2.1: First, we show that $D_0(\theta^o)$ is an empty set. If not, assume that there is a point $(z^*, X^*)$ in set $D_0(\theta^o)$. By definition (19), both points $(z^*, X^*)$ and $(-z^* - 2X^*\theta^o, X^*)$ are in set $int(S_{z,X})$. By Assumptions 1, 2, and equations (10)-(12), we have function

$$d_0(\theta^o; z^*, X^*) = (-1)^j [f_\xi (-z^* - X^*\theta^o | X = X^*) - f_\xi (z^* + X^*\theta^o | X = X^*)] = 0,$$

which is a contradiction with definition (19).
Next, we prove that $P[ (z^*, X^*) \in D_0(\theta) ] > 0$ for any $\theta \neq \theta^0$, where $\theta \in \Theta$ and parameter space $\Theta$ satisfies Assumption 3. Denote the set $\mathcal{X}(\theta) \equiv \{ X^* \in S_X | X^*(\theta - \theta^0) \neq 0 \}$, which is a collection of covariate values at which $X\theta \neq X\theta^0$. By Assumption 4(a) and the fact $\theta - \theta^0 \neq 0$, $\mathcal{X}(\theta)$ is a subset in the support of $S_X$ with positive measure, that is,

$$P[ X^* \in \mathcal{X}(\theta) ] > 0. $$ (A1)

Recall that we use $X_c$ and $X_d$, respectively, to denote the continuous and discrete covariates in $X$. Similar to (18), we define the interior of the support of $X$ as $int(S_X) \equiv \{(X_c^*, X_d^*) \in S_{(X_c,X_d)} | X_c^* \in int(S_{X_c}(X_d^*)), X_d^* \in S_{X_d} \}$. Define

$$\tilde{S}_{(z,X)}(\theta) \equiv \left\{ (z^*, X^*) \in S_{(z,X)} \mid z^* \in \tilde{S}_z(X^*), X^* \in \mathcal{X}(\theta) \cap int(S_X) \right\} ,$$ (A2)

where $\tilde{S}_z(X^*)$ satisfies Assumption 4(c). By construction, set $\tilde{S}_{(z,X)}(\theta)$ is a Lebesgue measurable subset of $int(S_{(z,X)})$. Next we construct a subset in the support of covariates $(z, X)$ as follows:

$$\tilde{D}_0(\theta) \equiv \left\{ (z^*, X^*) \in \tilde{S}_{(z,X)}(\theta) \mid d_0(\theta; z^*, X^*) \neq 0 \right\}$$ (A3)

which is also a subset of $D_0(\theta)$ because $(-z^* - 2X^*\theta, X^*) \in int(S_{(z,X)})$ for any $(z^*, X^*) \in \tilde{S}_{(z,X)}(\theta)$. Under Assumptions 1-4 and 5(a), both sets $D_0(\theta)$ and $\tilde{D}_0(\theta)$ are Lebesgue measurable.

Theorem 2.1 is proved if we show $P[ (z^*, X^*) \in \tilde{D}_0(\theta) ] > 0$ since $\tilde{D}_0(\theta) \subseteq D_0(\theta)$. Now

$$P \left[ (z^*, X^*) \in \tilde{S}_{(z,X)}(\theta) \right] = P \left[ X^* \in \mathcal{X}(\theta) \cap int(S_X) \right] \times P \left[ z^* \in \tilde{S}_z(X^*) \mid X^* \in \mathcal{X}(\theta) \cap int(S_X) \right] ,$$ (A4)

where the first probability on the right of equation (A4) is positive by (A1), and the second is positive by Assumption 4(c). Under Assumptions 1, 2, and equations (10)-(12), we have

$$d_0(\theta; z^*, X^*) = (-1)^t \left[ f_\varepsilon (-z^* - X^*\theta^0 | X = X^*) - f_\varepsilon (z^* + 2X^*\theta - X^*\theta^0 | X = X^*) \right]$$
for every \((z^*, X^*) \in \tilde{S}_{(z,X)}(\theta)\). Define \(r = 2X^*(\theta - \theta^\circ)\). Then \(r \neq 0_J\) because \(X^* \in \mathcal{X}(\theta)\). We can write \(z^* + 2X^*\theta - X^*\theta^\circ = r + z^* + X^*\theta^\circ\) so function

\[
d_0(\theta; z^*, X^*) = (-1)^J [f_\varepsilon (-z^* - X^*\theta^\circ | X = X^*) - f_\varepsilon (r + z^* + X^*\theta^\circ | X = X^*)]. \tag{A5}
\]

We claim that

\[
P \left( d_0(\theta; z^*, X^*) \neq 0 \mid (z^*, X^*) \in \tilde{S}_{(z,X)}(\theta) \right) > 0. \tag{A6}
\]

If (A6) is not true, then for almost every \((z^*, X^*) \in \tilde{S}_{(z,X)}(\theta)\) we get \(d_0(\theta; z^*, X^*) = 0\), which implies that \(f_\varepsilon (t | X = X^*) = f_\varepsilon (r - t | X = X^*)\) for every \(t \in \tilde{S}_\varepsilon(X^*)\) and \(r - t \in \tilde{S}_\varepsilon(X^*)\) by (A5) and Assumption 5(a). This is possible only if \(r = 0_J\) by Assumption 5(b), which contradicts \(r \neq 0_J\). We have therefore proved that \(P[(z^*, X^*) \in \tilde{D}_0(\theta)] > 0\) by (A3), (A4), and (A6).

Q.E.D.
References


*Econometrica*, 54, 1435-1460.


Supplemental Appendix to: Semiparametric Identification and Estimation of Multinomial Discrete Choice Models using Error Symmetry*

Arthur Lewbel†  Jin Yan‡  Yu Zhou§

Original February 2019, revised January 2021

S.A  Proofs Regarding Estimation

In Appendix S.A, we provide the proofs of Theorems 3.1-3.3 in Section 3 and their related lemmas. Specifically, Section S.A.1 provides the proof of Theorem 3.1 on the population sample objective function; Section S.A.2 collects preliminary lemmas needed for the asymptotic properties of the MD estimator defined in Section 3.2; Section S.A.3 provides the proofs of Theorem 3.2, the consistency of the MD estimator; and Section S.A.4 gives the proofs of Theorem 3.3, the asymptotic linearity and normality of the estimator and related lemmas. Throughout this appendix, we use the same notations and acronyms defined in the main text.

*We thank Songnian Chen, Jeremy Fox, Bruce Hansen, Shakeeb Khan, Brendan Kline, Rosa Matzkin, Byoung Park, Jack Porter, Andres Santos, Xiaoxia Shi, Matthew Shum, Christopher Taber, Yuanyuan Wan and Hanghui Zhang for valuable comments and discussions and seminar participants at Xiamen University, Academia Sinica, Renmin University of China, the 2019 Asia Meeting, China Meeting, and Australasian Meeting of the Econometric Society, and the 29th Annual Meeting of the Midwest Econometrics Group. Funding support from the DVPS of the Chinese University of Hong Kong is greatly appreciated. All errors are ours.

†Department of Economics, Boston College. E-mail: lewbel@bc.edu.
‡Department of Economics, The Chinese University of Hong Kong, Hong Kong. E-mail: jyan@cuhk.edu.hk.
§Department of Economics, School of Economics, Fudan University; Shanghai; email: yuzhou@fudan.edu.cn
S.A.1 Proof of the Population Objective Function

Proof of Theorem 3.1: Part (i) can be shown directly from the quadratic form of the population objective function. We will explicitly prove that Part (ii) holds. To show the existence of a minimizer, recall the population objective function

\[ Q_j(\theta) = \frac{1}{2} E \left[ \sigma_j (z_n^{(j)}, X_n^{(j)}) d_j (\theta; z_n^{(j)}, X_n^{(j)}) \right]^2 \]

\[ = \frac{1}{2} E \left\{ E \left[ \sigma_j^2 (z_n^{(j)}, X_n^{(j)}) d_j^2 (\theta; z_n^{(j)}, X_n^{(j)}) \right] \right\} \]

From the main identification restriction we discuss in Section 2, we have

\[ E \left[ \sigma_j^2 (z_n^{(j)}, X_n^{(j)}) d_j^2 (\theta; z_n^{(j)}, X_n^{(j)}) \right] = 0 \]  \hspace{1cm} (S.A.1)

when \( \theta = \theta^o \). The equality in (S.A.1) holds because, conditional on \( X_n^{(j)} \), when \( \sigma_j (z_n^{(j)}, X_n^{(j)}) > 0 \), \( d_j (\theta; z_n^{(j)}, X_n^{(j)}) = 0 \); and in addition, when \( \sigma_j (z_n^{(j)}, X_n^{(j)}) = 0 \), the product term in the expectation is also equal to zero. Combining these parts gives the desired existence.

To show the uniqueness, consider any \( \theta \) in the parameter space such that \( \theta \neq \theta^o \). We have

\[ Q_j(\theta) - Q_j(\theta^o) \]  \hspace{1cm} (S.A.2)

\[ = \frac{1}{2} E \left[ \sigma_j (z_n^{(j)}, X_n^{(j)}) d_j (\theta; z_n^{(j)}, X_n^{(j)}) \right]^2 - \frac{1}{2} E \left[ \sigma_j (z_n^{(j)}, X_n^{(j)}) d_j (\theta^o; z_n^{(j)}, X_n^{(j)}) \right]^2 \]

\[ = \frac{1}{2} E \left[ \sigma_j^2 (z_n^{(j)}, X_n^{(j)}) \left( d_j (\theta; z_n^{(j)}, X_n^{(j)}) - d_j (\theta^o; z_n^{(j)}, X_n^{(j)}) \right) \right]^2 \]

\[ + E \left[ \sigma_j^2 (z_n^{(j)}, X_n^{(j)}) \right] d_j (\theta^o; z_n^{(j)}, X_n^{(j)}) \left( d_j (\theta; z_n^{(j)}, X_n^{(j)}) - d_j (\theta^o; z_n^{(j)}, X_n^{(j)}) \right) \]

The last inequality in (S.A.2) holds because the first term on its right-hand side is strictly positive, as there exists some \( z_n^{(j)}, X_n^{(j)} \) such that \( \sigma_j (z_n^{(j)}, X_n^{(j)}) > 0 \) and \( d_j (\theta; z_n^{(j)}, X_n^{(j)}) - d_j (\theta^o; z_n^{(j)}, X_n^{(j)}) \neq 0 \) when \( \theta \neq \theta^o \) by Theorem 2.1 and the identification results in Section 2.5; and the second term equals to zero since \( d_j (\theta^o; z_n^{(j)}, X_n^{(j)}) = 0 \). Q.E.D.

2
S.A.2 Proofs of Some Lemmas for the Asymptotic Properties

Below we first derive some lemmas based on the Höfding decomposition for the asymptotic properties.

**Lemma S.A.1** (Lemma 3.1 in Powell et al. (1989) and Lemma D.1 in Chen et al. (2016)).

For an i.i.d. sequence of random variables, \( \{\omega_m, m = 1, ..., N\} \), define a general second-order U-statistic of the form

\[
U_N = \frac{1}{N(N-1)} \sum_{m=1, m\neq n}^{N} \sum_{n=1}^{N} \psi_N (\omega_m, \omega_n)
\]

Define

\[
\hat{U}_N = \mu_N + \frac{1}{N} \sum_{m=1}^{N} (r_{N1} (\omega_m) - \mu_N) + \frac{1}{N} \sum_{n=1}^{N} (r_{N2} (\omega_n) - \mu_N)
\]

where \( r_{N1} (\omega_m) \equiv E [\psi_N (\omega_m, \omega_n) | \omega_m], r_{N2} (\omega_n) \equiv E [\psi_N (\omega_m, \omega_n) | \omega_n], \) and \( \mu_N \equiv E [\psi_N (\omega_m, \omega_n)] = E [r_{N1} (\omega_m)] = E [r_{N2} (\omega_n)]. \) If \( E \|\psi_N (\omega_m, \omega_n)\|^2 = o (N) \), then \( U_N = \hat{U}_N + o_p (N^{-1/2}) \).

**Lemma S.A.2** Under Assumptions 8-11,

\[
\sup_{(z_n^{(j)}, x_n^{(j)}) \in \mathcal{S}_{T_{z_n^{(j)}, x_n^{(j)}}}^{T_{r}}} \left| \hat{f}_j \left( z_n^{(j)}, x_n^{(j)} \right) - f_j \left( z_n^{(j)}, x_n^{(j)} \right) \right| = O_p \left( \sqrt{\frac{\ln N}{Nh_N^{r+J_q}}} + h_N^r \right) = o_p \left( N^{-1/4} \right)
\]

\[
\sup_{(z_n^{(j)}, x_n^{(j)}) \in \mathcal{S}_{T_{z_n^{(j)}, x_n^{(j)}}}^{T_{r}}} \left| \hat{g}_j \left( z_n^{(j)}, x_n^{(j)} \right) - g_j \left( z_n^{(j)}, x_n^{(j)} \right) \right| = O_p \left( \sqrt{\frac{\ln N}{Nh_N^{r+J_q}}} + h_N^r \right) = o_p \left( N^{-1/4} \right)
\]

\[
\sup_{(z_n^{(j)}, x_n^{(j)}) \in \mathcal{S}_{T_{z_n^{(j)}, x_n^{(j)}}}^{T_{r}}} \left| \hat{\varphi}_j \left( z_n^{(j)}, x_n^{(j)} \right) - \varphi_j \left( z_n^{(j)}, x_n^{(j)} \right) \right| = O_p \left( \sqrt{\frac{\ln N}{Nh_N^{r+J_q}}} + h_N^r \right) = o_p \left( N^{-1/4} \right)
\]

**Proof of Lemma S.A.2:** The proofs for three terms are similar. We will focus on the proof for \( \hat{g}_j \left( z_n^{(j)}, x_n^{(j)} \right) \). Other terms can be done in a similar fashion. First, by the fact that the outcome come variables By the fact that the outcome variables are binary and function \( f_j \) is
bounded away from zero, applying the results of Lemma B.1 and Lemma B.2 in Newey (1994) gives the first equality in each equation. Second, the second equality follows from Assumption 10 using Lemma 8.10 in Newey and McFadden (1994). \textit{Q.E.D.}

Define $j^t_j(z^{(j)}, X^{(j)}) = \partial^t \hat{f}_j / \partial z_{1,t} \cdots \partial z_{t,t}$ be the derivative with respect to $z^{(j)}_{(t)}$, where $z^{(j)}_{(t)} = (z^{(j)}_{1,t}, \ldots, \partial z^{(j)}_{t,t})$ be any t-element of $z^{(j)}$. Similarly, we can define $f^{(t)}_j(z^{(j)}, X^{(j)})$, $g^{(t)}_j(z^{(j)}, X^{(j)})$, $\phi^{(t)}_j(z^{(j)}, X^{(j)})$ and $\varphi^{(t)}_j(z^{(j)}, X^{(j)})$.

**Lemma S.A.3** Under Assumptions 8-11, for $t = 1, \ldots, J$,

$$
\sup_{(z^{(j)}_{n}, X^{(j)}_{n}) \in S^{Tr}_{(z^{(j)}, X^{(j)})}} \quad \left| \hat{j}^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) - f^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) \right| = O_p \left( \sqrt{\frac{\ln N}{Nh^2_N + Jq} + h_N^s} \right) = o_p \left( N^{-1/4} \right)
$$

$$
\sup_{(z^{(j)}_{n}, X^{(j)}_{n}) \in S^{Tr}_{(z^{(j)}, X^{(j)})}} \quad \left| \hat{g}^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) - g^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) \right| = O_p \left( \sqrt{\frac{\ln N}{Nh^2_N + Jq} + h_N^s} \right) = o_p \left( N^{-1/4} \right)
$$

$$
\sup_{(z^{(j)}_{n}, X^{(j)}_{n}) \in S^{Tr}_{(z^{(j)}, X^{(j)})}} \quad \left| \hat{\phi}^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) - \phi^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) \right| = O_p \left( \sqrt{\frac{\ln N}{Nh^2_N + Jq} + h_N^s} \right) = o_p \left( N^{-1/4} \right)
$$

**Proof of Lemma S.A.3:** The proof follows the same method used in Lemma S.A.2. \textit{Q.E.D.}

**Lemma S.A.4** Under Assumptions 8-11, for $t = 1, \ldots, J$,

$$
\sup_{(z^{(j)}_{n}, X^{(j)}_{n}) \in S^{Tr}_{(z^{(j)}, X^{(j)})}} \quad \left\| \nabla_{\theta} \left( j^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) \right) - \nabla_{\theta} \left( f^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) \right) \right\|
$$

$$
= O_p \left( \sqrt{\frac{\ln N}{Nh^2_N + Jq} + h_N^s} \right) = o_p \left( N^{-1/4} \right)
$$

$$
\sup_{(z^{(j)}_{n}, X^{(j)}_{n}) \in S^{Tr}_{(z^{(j)}, X^{(j)})}} \quad \left\| \nabla_{\theta} \left( g^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) \right) - \nabla_{\theta} \left( g^{(t)}_j \left( z^{(j)}_{n}, X^{(j)}_{n} \right) \right) \right\|
$$

$$
= O_p \left( \sqrt{\frac{\ln N}{Nh^2_N + Jq} + h_N^s} \right) = o_p \left( N^{-1/4} \right)
$$
\[
\begin{align*}
\sup_{(z_n^{(j)}, X_n^{(j)}) \in \mathcal{S}_r^{TP}} \left\| \nabla_{\theta} \left( \varphi_j^{(t)} \left( z_n^{(j)}, X_n^{(j)} \right) \right) - \nabla_{\theta} \left( \varphi_j^{(t)} \left( z_n^{(j)}, X_n^{(j)} \right) \right) \right\| \\
= O_p \left( \sqrt{\frac{\ln N}{Nh_N^{J+2(t+1)+Jq}} + h_N^*} \right) = o_p \left( N^{-1/4} \right)
\end{align*}
\]

**Proof of Lemma S.A.4:** The proof follows the same method used in Lemma S.A.2. Q.E.D.

**S.A.3 Consistency of the MD Estimator**

**Proof of Theorem 3.2:** We apply Theorem 2.1 of Newey and McFadden (1994) to show the consistency of the MD estimator. Theorem 2.1 in Newey and McFadden (1994) requires the following fours conditions: (1) the population objective function \( Q_j(\theta) \) is uniquely minimized at \( \theta^0 \in \Theta \); (2) the parameter space \( \Theta \) is compact; (3) the population objective function \( Q_j(\theta) \) is continuous; and (4) the sample objective function \( Q_N(j)(\theta) \) converges uniformly in probability to \( Q_j(\theta) \) over the parameter space.

Our Theorem 3.1 directly implies Condition (1). Condition (2) follows from Assumption 3. Condition (3) is from the continuity of our population objective function. Below, we show Condition (4) following Hong and Tamer (2003). We first introduce an infeasible sample objective function \( \bar{Q}_N(j)(\theta) \), defined as

\[
\bar{Q}_N(j)(\theta) = \frac{1}{2N} \sum_{n=1}^{N} \left[ \tau_j \left( z_n^{(j)}, X_n^{(j)} \right) d_j \left( \theta; z_n^{(j)}, X_n^{(j)} \right) \right]^2.
\]

Following the triangle inequality, we have

\[
|Q_N(j)(\theta) - Q_j(\theta)| \leq |Q_N(j)(\theta) - \bar{Q}_N(j)(\theta)| + |\bar{Q}_N(j)(\theta) - Q_j(\theta)|.
\]

(S.A.3)

Then, it is sufficient to show that the two terms on the right side of (S.A.3) go to zero uniformly, that is, (i) \( \sup_{\theta \in \Theta} |Q_N(j)(\theta) - \bar{Q}_N(j)(\theta)| = o_p(1) \) and (ii) \( \sup_{\theta \in \Theta} |\bar{Q}_N(j)(\theta) - Q_j(\theta)| = o_p(1) \).
For Part (i), we observe that

\[
\sup_{\theta \in \Theta} \left| Q_{Nj}(\theta) - \tilde{Q}_{Nj}(\theta) \right| \quad (\text{S.A.4})
\]

\[
= \sup_{\theta \in \Theta} \left| \frac{1}{2N} \sum_{n=1}^{N} \left[ \tau_j^2 \left( z_n^{(j)}, X_n^{(j)} \right) \left\{ \hat{d}_{j,-n} \left( \theta; z_n^{(j)}, X_n^{(j)} \right) + d_j \left( \theta; z_n^{(j)}, X_n^{(j)} \right) \right\} \right] \right|
\]

\[
= \sup_{\theta \in \Theta} \left| \frac{1}{2N} \sum_{n=1}^{N} \tau_j^2 \left( z_n^{(j)}, X_n^{(j)} \right) \left\{ \hat{d}_{j,-n} \left( \theta; z_n^{(j)}, X_n^{(j)} \right) - d_j \left( \theta; z_n^{(j)}, X_n^{(j)} \right) \right\} \right|
\]

\[
\leq C \sup_{\theta \in \Theta} \sup_{(z_n^{(j)}, X_n^{(j)}) \in \mathcal{S}_T^r} \left| \hat{d}_{j,-n} \left( \theta; z_n^{(j)}, X_n^{(j)} \right) - d_j \left( \theta; z_n^{(j)}, X_n^{(j)} \right) \right| = o_p(1).
\]

The first equality in (S.A.4) follows from definition and direct calculation. The second equality holds by factorization. The next inequality is satisfied by the fact that functions \(\tau_j\) and \(d_j\) are bounded. The last equality follows the fact that

\[
\sup_{\theta \in \Theta} \sup_{(z_n^{(j)}, X_n^{(j)}) \in \mathcal{S}_T^r} \left| \hat{d}_{j,-n} \left( \theta; z_n^{(j)}, X_n^{(j)} \right) - d_j \left( \theta; z_n^{(j)}, X_n^{(j)} \right) \right|
\]

is bounded by the product of a constant and the derivative functions shown by Lemma S.A.3.

Part (ii) holds by showing pointwise convergence and stochastic equicontinuity. By the Law of Large Numbers (LLN), we can directly obtain the pointwise convergence of \(\tilde{Q}_{Nj}(\theta)\) to \(Q_j(\theta)\). Next we can conclude the uniformity by showing stochastic equicontinuity, that is,

\[
\sup_{\theta^{(1)}, \theta^{(2)} \in \Theta, \|\theta^{(1)} - \theta^{(2)}\| \leq \delta} \left| \tilde{Q}_{Nj}(\theta^{(1)}) - \tilde{Q}_{Nj}(\theta^{(2)}) \right| = o_p(1).
\]

Following Andrews (1994), the stochastic equicontinuity can be shown by verifying that \(\tilde{Q}_{Nj}(\theta)\) is the type II class of function, satisfying the Lipschitz condition

\[
\left| \tilde{Q}_{Nj}(\theta^{(1)}) - \tilde{Q}_{Nj}(\theta^{(2)}) \right| \leq C\|\theta^{(1)} - \theta^{(2)}\|. \quad \text{We verify that this holds from the continuity of the quadratic form of the objective}
\]
function and the continuity of the kernel derivative functions with bounded second derivatives.

Q.E.D.

S.A.4 Asymptotic Linearity and Normality of the MD Estimator

In this section, we first show the lemmas that contribute to the proof of Theorem 3.3.

Lemma S.A.5 Under Assumptions 1-12, \( H_{Nj} \left( \hat{\theta} \right) \to_p H_j \), where

\[
H_j = E \left\{ \tau_j^2 \left( z_n^{(j)} , X_n^{(j)} \right) \nabla_{\theta} d_j \left( \theta^o ; z_n^{(j)} , X_n^{(j)} \right) \left[ \nabla_{\theta} d_j \left( \theta^o ; z_n^{(j)} , X_n^{(j)} \right) \right]' \right\}
\]

Proof of Lemma S.A.5: To show the desired result, we first show that the following results hold:

(i) \( H_{Nj} \left( \hat{\theta} \right) = H_{Nj,1} \left( \hat{\theta} \right) + H_{Nj,2} \left( \hat{\theta} \right) \), where

\[
H_{Nj,1} \left( \hat{\theta} \right) = \frac{1}{N} \sum_{n=1}^{N} \tau_j^2 \left( z_n^{(j)} , X_n^{(j)} \right) \hat{d}_{j,-n} \left( \hat{\theta} ; z_n^{(j)} , X_n^{(j)} \right) \left[ \nabla_{\theta} \hat{d}_{j,-n} \left( \hat{\theta} ; z_n^{(j)} , X_n^{(j)} \right) \right]
\]

\[
H_{Nj,2} \left( \hat{\theta} \right) = \frac{1}{N} \sum_{n=1}^{N} \tau_j^2 \left( z_n^{(j)} , X_n^{(j)} \right) \nabla_{\theta} \hat{d}_{j,-n} \left( \hat{\theta} ; z_n^{(j)} , X_n^{(j)} \right) \left[ \nabla_{\theta} \hat{d}_{j,-n} \left( \hat{\theta} ; z_n^{(j)} , X_n^{(j)} \right) \right]'
\]

(ii) \( H_{Nj,1} \left( \hat{\theta} \right) = o_p \left( 1 \right) \), and (iii) \( H_{Nj,2} \left( \hat{\theta} \right) \to_p H_j \).

The decomposition in Part (i) follows from direct calculation. For Part (ii), observe that

\[
H_{Nj,1} \left( \hat{\theta} \right) = \left[ H_{Nj,1} \left( \hat{\theta} \right) - H_{Nj,1} \left( \theta^o \right) \right] + H_{Nj,1} \left( \theta^o \right) = o_p \left( 1 \right)
\]

Given that \( \hat{\theta} \) lies between \( \theta^o \) and \( \hat{\theta} \), we get that \( \hat{\theta} \) is uniformly consistent, and by applying the Delta method for the continuity of the choice probability, we obtain that \( H_{Nj,1} \left( \hat{\theta} \right) - H_{Nj,1} \left( \theta^o \right) = o_p \left( 1 \right) \). Next, \( H_{Nj,1} \left( \theta^o \right) = o_p \left( 1 \right) \) can be directly shown by applying the Markov Inequality, using the fact that

\[
\tau_j^2 \left( z_n^{(j)} , X_n^{(j)} \right) d_j \left( \theta^o ; z_n^{(j)} , X_n^{(j)} \right) \left[ \nabla_{\theta} d_j \left( \theta^o ; z_n^{(j)} , X_n^{(j)} \right) \right]' = 0_{q \times 1}
\]

7
and

$$
\sup_{\theta \in \Theta} \sup_{(z^{(j)}_n, X^{(j)}_n) \in S^{T_j}_N \times (z^{(j)}, X^{(j)})} \left| \hat{d}_{j,-n} \left( \theta; z^{(j)}_n, X^{(j)}_n \right) - d_j \left( \theta; z^{(j)}_n, X^{(j)}_n \right) \right|
$$

is bounded by the product of a constant and the derivative functions shown by Lemmas S.A.2 and S.A.3.

For Part (iii), define

$$
H_j (\theta) = E \left\{ \tau^2_j \left( z^{(j)}_n, X^{(j)}_n \right) \nabla_\theta d_j \left( \theta; z^{(j)}_n, X^{(j)}_n \right) \left[ \nabla_\theta d_j \left( \theta; z^{(j)}_n, X^{(j)}_n \right) \right] \right\}
$$

and we have

$$
H_{Nj,2} \left( \hat{\theta} \right) - H_j (\theta^o) = H_{Nj,2} \left( \hat{\theta} \right) - H_j (\theta^o) = \left[ H_{Nj,2} \left( \hat{\theta} \right) - H_j (\theta) \right] + \left[ H_j (\hat{\theta}) - H_j (\theta^o) \right].
$$

By the triangle inequality theorem, we have that

$$
\left\| H_{Nj,2} \left( \hat{\theta} \right) - H_j (\theta^o) \right\| \leq \left\| H_{Nj,2} \left( \hat{\theta} \right) - H_j (\theta) \right\| + \left\| H_j (\hat{\theta}) - H_j (\theta^o) \right\|.
$$

The desired results then follow from the strong LLN \( \sup_{\theta \in \Theta} \left\| H_{Nj,2} \left( \hat{\theta} \right) - H_j (\theta) \right\| \to_p 0_{q \times q} \)

and \( \hat{\theta} \) is a uniformly consistent estimator of \( \theta^o \). \( Q.E.D. \)

To analyze the properties of the numerator of the MD estimator, below we decompose it into two terms by adding and subtracting \( \nabla_\theta d_j \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \) in the square brackets,

$$
q_{Nj} (\theta^o) = \frac{1}{N} \sum_{n=1}^{N} \tau^2_j \left( z^{(j)}_n, X^{(j)}_n \right) \hat{d}_{j,-n} \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \left[ \nabla_\theta \hat{d}_{j,-n} \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \right] \tag{S.A.5}
$$

$$
= \frac{1}{N} \sum_{n=1}^{N} \tau^2_j \left( z^{(j)}_n, X^{(j)}_n \right) \hat{d}_{j,-n} \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \left[ \nabla_\theta d_j \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \right]
$$

$$
+ \frac{1}{N} \sum_{n=1}^{N} \tau^2_j \left( z^{(j)}_n, X^{(j)}_n \right) \hat{d}_{j,-n} \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \left[ \nabla_\theta \hat{d}_{j,-n} \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) - \nabla_\theta d_j \left( \theta^o; z^{(j)}_n, X^{(j)}_n \right) \right]
$$

$$
\equiv q_{Nj,1} (\theta^o) + q_{Nj,2} (\theta^o).
$$
We will show that the term \( q_{N,j,1}(\theta^o) \) on the right side of (S.A.5) contributes to the asymptotic distribution while the term \( q_{N,j,2}(\theta^o) \) is asymptotically negligible.

**Lemma S.A.6** Under Assumptions 1-11, we have

\[
q_{N,j,2}(\theta^o) = \frac{1}{N} \sum_{n=1}^{N} \tau_j^2 \left( z_n^{(j)}, X_n^{(j)} \right) \hat{d}_{j,-n} \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) - \nabla_{\theta} \hat{d}_{j,-n} \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) - \nabla_{\theta} d_j \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) = o_p \left( N^{-1/2} \right)
\]

Proof of Lemma S.A.6: Note that

\[
\hat{d}_{j,-n} \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) = \left[ \varphi_{j,o,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - \varphi_{j,o}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right] - \left[ \varphi_{j,cs,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - \varphi_{j,cs}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right] = \hat{d}_{j,o,-n} \left( z_n^{(j)}, X_n^{(j)} \right) - \hat{d}_{j,cs,-n} \left( z_n^{(j)}, X_n^{(j)} \right),
\]

where the first equality in (S.A.6) holds by \( d_j \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) = 0 \) and the second equality follows the definitions of \( \hat{d}_{j,-n} \) and \( d_j \) in Section 3 of the main text.

Next we calculate

\[
q_{N,j,2}(\theta^o) = \frac{1}{N} \sum_{n=1}^{N} \tau_j^2 \left( z_n^{(j)}, X_n^{(j)} \right) \left[ \varphi_{j,o,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - \varphi_{j,o}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right] + \varphi_{j,cs,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - \varphi_{j,cs}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right] \times \left[ \nabla_{\theta} \hat{d}_{j,-n} \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) - \nabla_{\theta} d_j \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) \right] = A_1 + A_2 + A_3 + A_4
\]
where

\[ A_1 = \frac{1}{N} \sum_{n=1}^{N} \left[ \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \varphi_{j,o=1-n}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \varphi_{j,o}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \right] \]

\times \left[ \sum_{j=1}^{J} \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \nabla_{\theta} \varphi_{j,o=1-n,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \nabla_{\theta} \varphi_{j,o,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \times x_{n,j}^{(j)} \right] \]

\[ A_2 = \frac{1}{N} \sum_{n=1}^{N} \left[ \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \varphi_{j,o=1-n}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \varphi_{j,o}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \right] \]

\times \left[ \sum_{j=1}^{J} \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \nabla_{\theta} \varphi_{j,cs=1-n,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \nabla_{\theta} \varphi_{j,cs,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \times x_{n,j}^{(j)} \right] \]

\[ A_3 = \frac{1}{N} \sum_{n=1}^{N} \left[ \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \varphi_{j,cs=1-n}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \varphi_{j,cs}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \right] \]

\times \left[ \sum_{j=1}^{J} \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \nabla_{\theta} \varphi_{j,cs=1-n,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \nabla_{\theta} \varphi_{j,cs,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \times x_{n,j}^{(j)} \right] \]

\[ A_4 = \frac{1}{N} \sum_{n=1}^{N} \left[ \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \varphi_{j,cs=1-n}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \varphi_{j,cs}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \right] \]

\times \left[ \sum_{j=1}^{J} \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \nabla_{\theta} \varphi_{j,cs=1-n,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \nabla_{\theta} \varphi_{j,cs,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \times x_{n,j}^{(j)} \right] \]

where \( j \) represents the choice of \( j \) product and \( (j) \) represents the derivatives with respect to \( j \) index. For \( A_1 \), we have

\[ \frac{1}{N} \sum_{n=1}^{N} \left[ \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \varphi_{j,o=1-n}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \varphi_{j,o}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \right] \]

\[ \leq \left( \sup_{(z_n^{(j)} , X_n^{(j)})} \left| \varphi_{j,o=1-n}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \varphi_{j,o}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right| \right)^2 \]

\[ = o_p \left( N^{-1/2} \right) \]

and in addition,

\[ \frac{1}{N} \sum_{n=1}^{N} \left[ \sum_{j=1}^{J} \tau_n \left( z_n^{(j)} , X_n^{(j)} \right) \left( \nabla_{\theta} \varphi_{j,o=1-n,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \nabla_{\theta} \varphi_{j,o,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right) \times x_{n,1}^{(j)} \right] \]

\[ \leq C_q \left( \sup_{(z_n^{(j)} , X_n^{(j)})} \left\| \nabla_{\theta} \varphi_{j,o=1-n,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) - \nabla_{\theta} \varphi_{j,o,j}^{(j)} (z_n^{(j)} , X_n^{(j)}) \right\| \right)^2 \]

\[ = o_p \left( N^{-1/2} \right) \]
where \( C_q \in \mathbb{R}^q \). Then by Cauchy-schwarz inequality, it follows that \( A_1 = o_p\left( N^{-1/2} \right) \). Similarly, we can show that \( A_2 = o_p\left( N^{-1/2} \right) \), \( A_3 = o_p\left( N^{-1/2} \right) \) and \( A_4 = o_p\left( N^{-1/2} \right) \). Combining all the results gives the desired results. Q.E.D.

Denote \( \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) = \tau_j^2 \left( z_n^{(j)}, X_n^{(j)} \right) \triangledown \theta d_j \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) \) for notational simplicity. By (S.A.6) we have

\[
q_{Nj,1}(\theta^o) = \frac{1}{N} \sum_{n=1}^{N} \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) d_{j,n} \left( \theta^o; z_n^{(j)}, X_n^{(j)} \right) \tag{S.A.8}
\]

\[
= \frac{1}{N} \sum_{n=1}^{N} \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left[ \hat{\phi}_{j,a,n} \left( z_n^{(j)}, X_n^{(j)} \right) - \hat{\phi}_{j,cs,n} \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \right].
\]

Lemma S.A.7 Under Assumptions 1-11,

\[
q_{Nj,1}(\theta^o) = \frac{1}{N(N-1)} \sum_{m=1}^{N} \sum_{n \neq m}^{N} \psi_N \left( \omega_m, \omega_n \right) + o_p\left( N^{-1/2} \right),
\]

where \( \psi_N \left( \omega_m, \omega_n \right) = \psi_{N,o} \left( \omega_m, \omega_n \right) - \psi_{N,cs} \left( \omega_m, \omega_n \right) \) with

\[
\psi_{N,o} \left( \omega_m, \omega_n \right) = \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left( y_{mj} - \varphi_{j,o} \left( z_n^{(j)}, X_n^{(j)} \right) \right)\
\times K_{h_s}^{(j)} \left( z_m^{(j)} - z_n^{(j)} \right) K_{h_X} \left( X_m^{(j)} - X_n^{(j)} \right) f_j^{-1} \left( z_n^{(j)}, X_n^{(j)} \right),
\]

\[
\psi_{N,cs} \left( \omega_m, \omega_n \right) = \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left( y_{mj} - \varphi_{j,cs} \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \right)\
\times K_{h_s}^{(j)} \left( z_m^{(j)} - \left( -z_n^{(j)} - 2X_n^{(j)} \theta^o \right) \right) K_{h_X} \left( X_m^{(j)} - X_n^{(j)} \right) f_j^{-1} \left( -z_n^{(j)} - 2X_n^{(j)} \theta^o, X_n^{(j)} \right).
\]

where \( K_{h_s}^{(j)} \left( z_m^{(j)} - \cdot \right) = \prod_{l=1}^{J} h_N^{-2} k_{l}^{(1)} \left( h_s^{-1} \left( z_{ml}^{(j)} - \cdot \right) \right) \) where \( k_{l}^{(1)} \) is the first derivative of kernel function.
Proof of Lemma S.A.7: We first observe that

\[ q_{Nj,1}(\theta^o) = \frac{1}{N} \sum_{n=1}^{N} \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \tilde{q}_{j,n} \left( \theta^o, z_n^{(j)}, X_n^{(j)} \right) \quad (S.A.9) \]

\[ \frac{1}{N} \sum_{n=1}^{N} \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left[ \hat{\varphi}_{j,o,n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - \hat{\varphi}_{j,o,n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right] \]

\[ \frac{1}{N} \sum_{n=1}^{N} \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left\{ \left[ \hat{\varphi}_{j,o,n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - \hat{\varphi}_{j,o,n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right] \right\} \]

\[ \frac{1}{N} \sum_{n=1}^{N} \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left\{ \left[ \hat{\varphi}_{j,o,n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - \hat{\varphi}_{j,o,n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right] \right\} + O(h^s) \]

The second, third and fourth equalities follows from adding and substracting terms. The last equality holds by the fact that

\[ \sup_{(z_n^{(j)}, X_n^{(j)}) \in S^T_{(z_n^{(j)}, X_n^{(j)})}} \left| E \left[ \hat{\varphi}_{j,o,n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \big| z_n^{(j)}, X_n^{(j)} \right] - \hat{\varphi}_{j,o}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right| = O(h^s) \]

\[ \sup_{(z_n^{(j)}, X_n^{(j)}) \in S^T_{(z_n^{(j)}, X_n^{(j)})}} \left| E \left[ \hat{\varphi}_{j,o,n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \big| z_n^{(j)}, X_n^{(j)} \right] - \hat{\varphi}_{j,o}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right| = O(h^s) \]
Next, to derive \( \hat{\varphi}_{j,o,-n}^{(j)} (z_n^{(j)}, X_n^{(j)}) - E \left[ \hat{\varphi}_{j,o,-n}^{(j)} (z_n^{(j)}, X_n^{(j)}) \mid z_n^{(j)}, X_n^{(j)} \right] \), we observe that

\[
\hat{\varphi}_{j,o,-n}^{(j)} (z_n^{(j)}, X_n^{(j)}, \theta^o) - E \left[ \hat{\varphi}_{j,o,-n}^{(j)} (z_n^{(j)}, X_n^{(j)}) \mid z_n^{(j)}, X_n^{(j)} \right] = f_j^{-1} \frac{1}{N} \sum_{m=1, m \neq n}^N y_{mj} K_h^{(j)} (z_m^{(j)} - z_n^{(j)}) K_h (X_m^{(j)} - X_n^{(j)}) \]

\[- E \left[ \hat{\varphi}_{j,o,-n}^{(j)} (z_n^{(j)}, X_n^{(j)}) \mid z_n^{(j)}, X_n^{(j)} \right] \]

\[= f_j^{-1} \frac{1}{N} \sum_{m=1, m \neq n}^N (y_{mj} - \hat{\varphi}_{j,o}^{(j)} (z_n^{(j)}, X_n^{(j)})) \times K_h^{(j)} (z_m^{(j)} - z_n^{(j)}) K_h (X_m^{(j)} - X_n^{(j)}) + R_{o,1} \]

where the second equality holds by expanding \( \hat{f}_j^{-1} \), and in addition \( R_{o,1} \) collects the higher order terms from the decomposition of \( \hat{f}_j^{-1} \). Note that \( R_{o,1} \) is bounded by the product of

\[
\sup_{(z_n^{(j)}, X_n^{(j)}) \in \mathcal{S}_{(z^{(j)}, X^{(j)})}^{T_r}} \left| E \left[ \hat{\varphi}_{j,o}^{(j)} (z_n^{(j)}, X_n^{(j)}) \mid z_n^{(j)}, X_n^{(j)} \right] - \varphi_{j,o}^{(j)} (z_n^{(j)}, X_n^{(j)}) \right|,
\]

and

\[
\sup_{(z_n^{(j)}, X_n^{(j)}) \in \mathcal{S}_{(z^{(j)}, X^{(j)})}^{T_r}} \left| \hat{f}_j (z_n^{(j)}, X_n^{(j)}) - f_j (z_n^{(j)}, X_n^{(j)}) \right|.
\]

Since each term is of order \( O_p (N^{-1/4}) \), thus \( R_{o,1} \) is of order \( O_p (N^{-1/2}) \). Denoting

\[
\psi_{N,o,1} (\omega_m, \omega_n) = \xi_j (z_n^{(j)}, X_n^{(j)}, \theta^o) \times f_j^{-1} \left( y_{mj} - \varphi_{j,o}^{(j)} (z_n^{(j)}, X_n^{(j)}) \right) K_h^{(j)} (z_m^{(j)} - z_n^{(j)}) K_h (X_m^{(j)} - X_n^{(j)})
\]

will give the first term of \( \psi_{N,1} (\omega_m, \omega_n) \).

\[^1 \hat{f}_j^{-1} = f_j^{-1} \left( 1 - f_j^{-1} \left( \hat{f}_j - f_j \right) + 2 f_j^{-2} \left( \hat{f}_j - f_j \right)^2 + o \left( \hat{f}_j - f_j \right)^2 \right) \]
In addition, to derive \( \hat{\varphi}_{j,cs,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - E \left[ \hat{\varphi}_{j,cs,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right| z_n^{(j)}, X_n^{(j)} \right] \), we observe that

\[
\hat{\varphi}_{j,cs,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) - E \left[ \hat{\varphi}_{j,cs,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right| z_n^{(j)}, X_n^{(j)} \right] = \hat{f}_j^{-1} y_{m_j} K_{h_x}^{(j)} \left( z_m^{(j)} - \left( -z_n^{(j)} - 2\theta^o X_n^{(j)} \right) \right) K_{h_x} \left( X_m^{(j)} - X_n^{(j)} \right)
\]

\[
E \left[ \hat{\varphi}_{j,cs,-n}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right| z_n^{(j)}, X_n^{(j)} \right] = \hat{f}_j^{-1} \frac{1}{N} \sum_{m=1, m \neq n}^N \left( y_{m_j} - \hat{\varphi}_{j,cs}^{(j)} \left( z_n^{(j)}, X_n^{(j)} \right) \right) \times K_{h_x}^{(j)} \left( z_m^{(j)} - \left( -z_n^{(j)} - 2\theta^o X_n^{(j)} \right) \right) K_{h_x} \left( X_m^{(j)} - X_n^{(j)} \right) + R_{cs,1}
\]

where the second equality holds by the same argument for \( \hat{f}_j^{-1} \), and \( R_{cs,1} \) collects the higher order terms from the decomposition of \( \hat{f}_j^{-1} \), with the order of \( O_p \left( N^{-1/2} \right) \), by the same argument as above. Denoting

\[
\psi_{N,cs,1} \left( \omega_m, \omega_n \right) = \xi_j \left( z_n^{(j)}, X_n^{(j)} \right)
\]

\[
\times \hat{f}_j^{-1} \left( y_{m_j} - \hat{\varphi}_{j,cs}^{(j)} \right) K_{h_x}^{(j)} \left( z_m^{(j)} - \left( -z_n^{(j)} - 2\theta^o X_n^{(j)} \right) \right) K_{h_x} \left( X_m^{(j)} - X_n^{(j)} \right)
\]

will give the second term of \( \psi_{N,1} \left( \omega_m, \omega_n \right) \).

Combining all the terms gives the desired results. Q.E.D.

**Lemma S.A.8** Under Assumptions 8-11,

\[
\frac{1}{N (N - 1)} \sum_{m=1}^N \sum_{n=1, n \neq m}^N \psi_N \left( \omega_m, \omega_n \right) = \frac{1}{N} \sum_{m=1}^N t_{m_j} + o_p \left( N^{-1/2} \right)
\]

and

\[
N^{-1/2} \sum_{m=1}^N t_{m_j} \rightarrow_d N \left( 0_q, \Omega_j \right),
\]

where \( t_{m_j} = (v_{m_j, o} - v_{m_j, cs}) \partial^j \xi_j \left( z_m^{(j)}, X_m^{(j)}, \theta^o \right) / \partial z_1^{(j)} \cdots \partial z_j^{(j)} \) and \( \Omega_j = E \left[ t_{m_j} t_{m_j}' \right] \).
Proof of Lemma S.A.8: We denote $U_n (\omega_m, \omega_n)$ as the second-order U-statistic and $\hat{U}_n (\omega_m, \omega_n)$ as the projection of the second-order U-statistic, which is given by

$$U_n (\omega_m, \omega_n) = \frac{1}{N(N-1)} \sum_{m=1, m \neq n}^{N} \sum_{n=1}^{N} \psi_N (\omega_m, \omega_n)$$

and

$$\hat{U}_n = E[\psi_N (\omega_m, \omega_n)] + \frac{1}{N} \sum_{m=1}^{N} (r_{N1} (\omega_m) - E[\psi_N (\omega_m, \omega_n)])$$

$$+ \frac{1}{N} \sum_{n=1}^{N} (r_{N2} (\omega_n) - E[\psi_N (\omega_m, \omega_n)]) ,$$

where $r_{N1} (\omega_m) = E[\psi_N (\omega_m, \omega_n) | \omega_m]$ and $r_{N2} (\omega_n) = E[\psi_N (\omega_m, \omega_n) | \omega_n]$. To apply this to Lemma S.A.1, we first show that $E \left[ \left\| \psi_N (\omega_m, \omega_n) \right\|^2 \right] = o(N)$, which is equivalent to showing that

$$E \left[ \left\| \psi_{N, o} (\omega_m, \omega_n) \right\|^2 \right] = o(N)$$

and $E \left[ \left\| \psi_{N, cs} (\omega_m, \omega_n) \right\|^2 \right] = o(N)$. Recall that $h_z = (h_N, \ldots, h_N)'$ and $h_X = (h_N, \ldots, h_N, \ldots, h_N)'$ in Assumption 10. Denote $u_{z}^{(j)} = h_{N}^{-1} (z_{n}^{(j)} - z_{n}^{(j)})$ and $u_{X}^{(j)} = h_{N}^{-1} \left[ \wedge \left( X_{m}^{(j)} - X_{n}^{(j)} \right) \right]$, where $\wedge$ is juxtaposing the consecutive rows of the matrix next to each other. In addition, define $\wedge$ as the inverse transformation (stacking the vector into a matrix) of $\wedge$. By direct calculation

$$E \left[ \left\| \psi_{N,o} (\omega_m, \omega_n) \right\|^2 \right]$$

\(= \int \left\| K_{h_{z}}^{(j)} \left( z_{m}^{(j)} - z_{n}^{(j)} \right) K_{h_{X}} \left( X_{m}^{(j)} - X_{n}^{(j)} \right) \right\|^2$$

$$\times \left[ \varphi_{j} \left( z_{m}^{(j)} , X_{m}^{(j)} \right) + \varphi_{j}^{2} \left( z_{n}^{(j)} , X_{n}^{(j)} \right) - 2\varphi_{j} \left( z_{m}^{(j)} , X_{m}^{(j)} \right) \varphi_{j} \left( z_{n}^{(j)} , X_{n}^{(j)} \right) \right]$$

$$\times f_{j} \left( z_{m}^{(j)} , X_{m}^{(j)} \right) f_{j} \left( z_{n}^{(j)} , X_{n}^{(j)} \right) f_{j}^{-1} \left( z_{m}^{(j)} , X_{m}^{(j)} \right) \xi_{j}^{2} \left( z_{n}^{(j)} , X_{n}^{(j)} , \theta' \right) dz_{m}^{(j)} dX_{m}^{(j)} dz_{n}^{(j)} dX_{n}^{(j)}$$

(S.A.10)
where the first equality in (S.A.10) follows from definitions; the second equality holds using a change of variables; and the third equality is satisfied by Assumptions 9 and 10. The desired result then follows from Assumption 10. Similarly, we can show

\[ E(h \kappa N; cs(m, n)) = o(N) \]

Next we show that the second term in \( \hat{U}_n \) contributes to the asymptotic linearity and normality, while the first and third terms are asymptotically negligible. In sum, we show that (i) \( E(\psi N(m, n)) = E[r_{N1}(m)] = E[r_{N2}(n)] = o(N^{-1/2}) \), (ii) \( \frac{1}{N} \sum_{n=1}^{N} (r_{N1}(m) - E[r_{N2}(n)]) = o_p(N^{-1/2}) \), and (iii) \( \frac{1}{N} \sum_{m=1}^{N} (r_{N1}(m) - E[r_{N1}(m)]) = N^{-1} \sum_{m=1}^{N} t_{mj}, \text{where } N^{-1/2} \sum_{m=1}^{N} t_{mj} \) is \( N(0_q, \Omega_j) \).

First, to show Part (i) holds, it is equivalent to show \( E(\psi N, o(m, n)) = o_p(N^{-1/2}) \) and \( E(\psi N, cs(m, n)) = o_p(N^{-1/2}) \).
\[ E \left[ \psi_{N,o} (\omega_m, \omega_n) \right] = E \left[ E \left[ \psi_{N,o} (\omega_m, \omega_n) \mid \omega_n \right] \right] \]

\[ = E \left[ \left. E \left[ \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left( y_{mj} - \varphi_{j,o} \left( z_n^{(j)}, X_n^{(j)} \right) \right) \right] \right| \omega_n \right] \]

\[ \times K_{h_x}^{(j)} \left( z_m^{(j)} - z_n^{(j)} \right) K_{h_x} \left( X_m^{(j)} - X_n^{(j)} \right) \times f_j^{-1} \left( z_n^{(j)}, X_n^{(j)} \mid \omega_n \right) \]

\[ = E \left[ \int \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left( \varphi_{j,o} \left( z_n^{(j)} + u_z^{(j)} h_z, X_n^{(j)} + \tilde{h} \left( u_X^{(j)} h_X \right) \right) - \varphi_{j,o} \left( z_n^{(j)}, X_n^{(j)} \right) \right) \right. \]

\[ \times \left. K_{h_x}^{(j)} \left( u_z^{(j)} \right) K_{h_x} \left( u_X^{(j)} \right) f_j^{-1} \left( z_n^{(j)}, X_n^{(j)} \right) \right] \]

\[ \times f_j^{-1} \left( z_n^{(j)} + u_z^{(j)} h_z, X_n^{(j)} + \tilde{h} \left( u_X^{(j)} h_X \right) \right) d u_z^{(j)} d u_X^{(j)} \mid \omega_n \]

\[ = -E \left[ \int \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \frac{\partial}{\partial u_z^{(j)}} \left( \varphi_{j,o} \left( z_n^{(j)} + u_z^{(j)} h_z, X_n^{(j)} + \tilde{h} \left( u_X^{(j)} h_X \right) \right) - \varphi_{j,o} \left( z_n^{(j)}, X_n^{(j)} \right) \right) \right. \]

\[ \times \left. K_{h_x} \left( u_z^{(j)} \right) K_{h_x} \left( u_X^{(j)} \right) f_j^{-1} \left( z_n^{(j)}, X_n^{(j)} \right) \right| \omega_n \]

\[ = O \left( h_N^s \right). \]

In addition, we can show that 
\[ E \left[ \psi_{N,cs} (\omega_m, \omega_n) \right] = E \left[ E \left[ \psi_{N,cs} (\omega_m, \omega_n) \mid \omega_n \right] \right] = O \left( h_N^s \right). \] Then it implies that

\[ E \left[ \psi_N (\omega_m, \omega_n) \right] = E \left[ \psi_{N,o} (\omega_m, \omega_n) \right] - E \left[ \psi_{N,cs} (\omega_m, \omega_n) \right] = O \left( h_N^s \right). \]
Second, to show Part (ii) holds, by direct calculation, we have

\[ r_{N,2,o}(\omega_n) = E\left[\psi_{N,o}(\omega_m,\omega_n) | \omega_n\right] \]  

\( (S.A.11) \)

\[
= E \left[ \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left( y_m - \varphi_j,o \left( z_n^{(j)}, X_n^{(j)} \right) \right) \right] \\
\times K_{h_x}^{(j)} \left( z_n^{(j)} - z_n^{(j)} \right) K_{h_x} \left( X_m^{(j)} - X_n^{(j)} \right) f_j^{-1} \left( z_n^{(j)}, X_n^{(j)} \right) | \omega_n] \\
= E \left[ \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left( \varphi_j,o \left( z_m^{(j)}, X_m^{(j)} \right) - \varphi_j,o \left( z_n^{(j)}, X_n^{(j)} \right) \right) \right] \\
\times K_{h_x}^{(j)} \left( z_n^{(j)} - z_n^{(j)} \right) K_{h_x} \left( X_m^{(j)} - X_n^{(j)} \right) f_j^{-1} \left( z_n^{(j)}, X_n^{(j)} \right) | \omega_n] \\
= \int \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \frac{\partial \left( \varphi_j,o \left( z_n^{(j)} + u_z^{(j)} h_z, X_n^{(j)} + \tilde{\lambda} \left( u_X^{(j)} h_X \right) \right) - \varphi_j,o \left( z_n^{(j)}, X_n^{(j)} \right) \right)}{\partial u_z^{(j)}} \\
\times K_{h_x} \left( u_z^{(j)} \right) K_{h_x} \left( u_X^{(j)} \right) f_j^{-1} \left( z_n^{(j)}, X_n^{(j)} \right) \\
\times f_j \left( z_n^{(j)} + u_z^{(j)} h_z, X_n^{(j)} + \tilde{\lambda} \left( u_X^{(j)} h_X \right) \right) du_z^{(j)} du_X^{(j)} \\
= O \left( h_N^{8} \right) = o \left( N^{-1/2} \right). 
\]

The last second equality follows from integration by parts and a Taylor expansion. We therefore get that

\[
\frac{1}{N} \sum_{n=1}^{N} \left( r_{N,2} \left( \omega_n \right) - E \left[ r_{N,2} \left( \omega_n \right) \right] \right) = o_p \left( N^{-1/2} \right). 
\]
To show Part (iii), we have

\[ r_{N1,o}(\omega_m) = E \left[ \psi_{N,o}(\omega_m, \omega_n) \mid \omega_m \right] \quad \text{(S.A.12)} \]

\[
= E \left[ \xi_j \left( z_n^{(j)}, X_n^{(j)}, \theta^o \right) \left( y_{mj} - \varphi_{j,o} \left( z_n^{(j)}, X_n^{(j)} \right) \right) \times K_{h_z}^{(j)} \left( z_m^{(j)} - z_n^{(j)} \right) \times K_{h_X} \left( X_m^{(j)} - X_n^{(j)} \right) \times f_j^{-1} \left( z_n^{(j)}, X_n^{(j)} \right) \mid \omega_m \right] 
\]

\[
= \int \xi_j \left( z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda} \left( u_X^{(j)} h_X \right) \right) \times \left( y_{mj} - \varphi_{j,o} \left( z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda} \left( u_X^{(j)} h_X \right) \right) \right) 
\times K_{h_z}^{(j)} \left( u_z^{(j)} \right) \times K_{h_X} \left( u_X^{(j)} \right) \times d u_z \times d u_X 
\]

\[
= \int \left[ \frac{\partial^j \xi_j \left( z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda} \left( u_X^{(j)} h_X \right) \right) \times \left( y_{mj} - \varphi_{j,o} \left( z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda} \left( u_X^{(j)} h_X \right) \right) \right)}{\partial z_1^{(j)} \cdots \partial z_f^{(j)}} \right] 
\times K_{h_z}^{(j)} \left( u_z^{(j)} \right) \times K_{h_X} \left( u_X^{(j)} \right) \times d u_z \times d u_X 
\]

\[
- y_{mj} \int \frac{\partial^j \xi_j \left( z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda} \left( u_X^{(j)} h_X \right) \right) \times \left( y_{mj} - \varphi_{j,o} \left( z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda} \left( u_X^{(j)} h_X \right) \right) \right)}{\partial z_1^{(j)} \cdots \partial z_f^{(j)}} \times K_{h_z} \left( u_z^{(j)} \right) \times K_{h_X} \left( u_X^{(j)} \right) \times d u_z \times d u_X 
\]

\[
= r_o (\omega_m) + \epsilon_{N,o} (\omega_m) 
\]
where

\[ r_o(\omega_m) = \int \frac{\partial^J \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o) \varphi_j(z_m^{(j)}, X_m^{(j)})}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} K_{h_z}(u_z^{(j)}) K_{h_X}(u_X^{(j)}) \, d u_z^{(j)} d u_X^{(j)} \quad (S.A.13) \]

\[ - y_{mj} \int \frac{\partial^J \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o)}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} K_{h_z}(u_z^{(j)}) K_{h_X}(u_X^{(j)}) \, d u_z^{(j)} d u_X^{(j)} \]

\[ = \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o) \frac{\partial^J \varphi_j, o(z_m^{(j)}, X_m^{(j)})}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \]

\[ - (y_{mj} - \varphi_j, o(z_m^{(j)}, X_m^{(j)})) \frac{\partial^J \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o)}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \]

and

\[ \varsigma_{N,o}(\omega_m) = \int \frac{\partial^J \xi_j(z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda}(u_X^{(j)} h_X), \theta^o) \varphi_j, o(z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda}(u_X^{(j)} h_X))}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \]

\[ (S.A.14) \]

\[ - \frac{\partial^J \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o) \varphi_j, o(z_m^{(j)}, X_m^{(j)})}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \]

\[ K_{h_z}(u_z^{(j)}) K_{h_X}(u_X^{(j)}) \, d u_z^{(j)} d u_X^{(j)} \]

\[ - y_{mj} \int \left[ \frac{\partial^J \xi_j(z_m^{(j)} - u_z^{(j)} h_z, X_m^{(j)} - \bar{\lambda}(u_X^{(j)} h_X), \theta^o)}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} - \frac{\partial^J \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o)}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \right] 

\times K_{h_z}(u_z^{(j)}) K_{h_X}(u_X^{(j)}) \, d u_z^{(j)} d u_X^{(j)} \]

Then it follows that

\[ \frac{1}{\sqrt{N}} \sum_{m=1}^{N} (r_{N1,o}(\omega_m) - E[r_{N1,o}(\omega_m)]) = \frac{1}{\sqrt{N}} \sum_{m=1}^{N} (r_o(\omega_m) - E[r_o(\omega_m)]) \]

\[ + \frac{1}{\sqrt{N}} \sum_{m=1}^{N} (\varsigma_{N,o}(\omega_m) - E[\varsigma_{N,o}(\omega_m)]) . \]

Then the limiting distribution of \( \frac{1}{\sqrt{N}} \sum_{m=1}^{N} (r_{N1,o}(\omega_m) - E[r_{N1,o}(\omega_m)]) \) is equivalent to the limiting distribution \( \frac{1}{\sqrt{N}} \sum_{m=1}^{N} (r_o(\omega_m) - E[r_o(\omega_m)]) \), provided that the second term converges in
probability to zero. Note that

\[
\frac{1}{\sqrt{N}} \sum_{m=1}^{N} (r_o(\omega_m) - E[r_o(\omega_m)]) \tag{S.A.15}
\]

\[
= \frac{1}{\sqrt{N}} \sum_{m=1}^{N} \left( \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o) \frac{\partial_j \varphi_{j,o}(z_m^{(j)}, X_m^{(j)})}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}} \right)
\]

\[
- E \left[ \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o) \varphi_{j,o}(z_m^{(j)}, X_m^{(j)}) \right]
\]

\[
- \frac{1}{\sqrt{N}} \sum_{m=1}^{N} \left( y_{mj} - \varphi_{j,o}(z_m^{(j)}, X_m^{(j)}) \right) \frac{\partial_j \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}}
\]

\[
+ E \left[ \left( y_{mj} - \varphi_{j,o}(z_m^{(j)}, X_m^{(j)}) \right) \frac{\partial_j \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}} \right]
\]

and

\[
\frac{1}{\sqrt{N}} \sum_{m=1}^{N} (r_{cs}(\omega_m) - E[r_{cs}(\omega_m)]) \tag{S.A.16}
\]

\[
= \frac{1}{\sqrt{N}} \sum_{m=1}^{N} \left( \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o) \frac{\partial_j \varphi_{j,cs}(z_m^{(j)}, X_m^{(j)}, \theta^o)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}} \right)
\]

\[
- E \left[ \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o) \varphi_{j,cs}(z_m^{(j)}, X_m^{(j)}, \theta^o) \right]
\]

\[
- \frac{1}{\sqrt{N}} \sum_{m=1}^{N} \left( y_{mj} - \varphi_{j,cs}(z_m^{(j)}, X_m^{(j)}, \theta^o) \right) \frac{\partial_j \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}}
\]

\[
+ E \left[ \left( y_{mj} - \varphi_{j,cs}(z_m^{(j)}, X_m^{(j)}, \theta^o) \right) \frac{\partial_j \xi_j(z_m^{(j)}, X_m^{(j)}, \theta^o)}{\partial z_1^{(j)} \ldots \partial z_j^{(j)}} \right]
\]
Then
\[
\frac{1}{\sqrt{N}} \sum_{m=1}^{N} (r(\omega_m) - E[r(\omega_m)])
= \frac{1}{\sqrt{N}} \sum_{m=1}^{N} (r_o(\omega_m) - E[r_o(\omega_m)]) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (r_{cs}(\omega_m) - E[r_{cs}(\omega_m)])
= -\frac{1}{\sqrt{N}} \sum_{m=1}^{N} \left( y_{mj} - \varphi_{j,o} \left( z_m^{(j)}, X_m^{(j)} \right) \right) \frac{\partial^J \xi_j}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \left( z_m^{(j)}, X_m^{(j)}, \theta^o \right)
+ \frac{1}{\sqrt{N}} \sum_{m=1}^{N} \left( y_{mj} - \varphi_{j,cs} \left( z_m^{(j)}, X_m^{(j)}, \theta^o \right) \right) \frac{\partial^J \xi_j}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \left( z_m^{(j)}, X_m^{(j)}, \theta^o \right)
= \frac{1}{\sqrt{N}} \sum_{m=1}^{N} t_{mj}.
\]

The first two terms in equation (S.A.15) cancel out with the first two terms in equation (S.A.16) by the identification equation. In addition, by Assumption 11 (similar to Assumption 3 in Powell et al. (1989)), this last term \( \frac{1}{\sqrt{N}} \sum_{m=1}^{N} (\varsigma_{N,o}(\omega_m) - E[\varsigma_{N,o}(\omega_m)]) \) has second moment, that is bounded by
\[4h_z^{2J+2Jq} \left\{ E \left[ (1 + |y| + \|z\|) m(z, \gamma) \right] \left[ \int \|u\| |K(u)| du \right]^2 \right\} = \mathcal{O} \left( h_z^{2J+2Jq} \right).\] So it will converge to zero in probability. Applying Linderberg-Feller Central Limit Theorem using Assumption 8 gives the desired results, where \( \Omega_j = E \left[ t_{mj} t_{mj}' \right]. \) Q.E.D.

**Proof of Theorem 3.3:** The limit distribution in Theorem 3.3 then follows from Lemmas S.A.5–S.A.8 and the non-singularity of \( H_j \) in Assumption 12 as well as the symmetry of the indices \( m \) and \( n. \) Q.E.D.

**Additional References**
