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

Arthur Lewbel† Jin Yan‡ Yu Zhou§

Original February 2019, revised December 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.

1 Introduction

Traditional multinomial choice models, such as multinomial logit (MNL) and multinomial probit (MNP), e.g., McFadden (1974), assume homoskedastic errors. However, in reality substantial

*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.

§Economics, New York University Shanghai; Shanghai. Email: amanda.yu.zhou@nyu.edu

unobserved heterogeneity is common, 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 propose a corresponding estimator, and show it's root N consistent and asymptotically normal.

Popular multinomial choice specifications that permit unobserved heterogeneity, such as Hausman and Wise (1978), McFadden and Train (2000), Train (2009), and the demand side of Berry, Levinsohn, and Pakes (1995), assume random coefficients. Our model nests random coefficient models as a special case (assuming they are symmetrically distributed), and so in particular nests the usual parametric assumption of normally distributed random coefficients.

Advantages of our model (and associated estimator) over standard multinomial choice with random coefficients include:

1. We don't need to specify which regressors have random coefficients.

2. The distribution of the random coefficients can depend on regressors.

3. We don't need to specify the functional form of the random coefficient distributions.

4. Our estimator remains numerically the same regardless of which regressors have random coefficients, and

5. Our estimator doesn't require numerical integration or deconvolution techniques.

Our key identifying assumption is error symmetry. We assume that the joint latent error distribution, conditional on covariates, is centrally symmetric.[1] Though error symmetry has not previously been used for identification and estimation of multinomial discrete choice models, it has been used in binary choice models. Manski (1988) shows that conditional symmetry in binary response models does not have identifying power beyond median independence. Chen, Khan,

[1]Many multivariate distributions are centrally symmetric. See, e.g., Serfling (2006). A partial list includes multivariate normal distributions, multivariate logistic, and the elliptically-contoured distribution as well as mean zero mixtures of multivariate normal distributions. MNP models with or without normal random coefficients have conditionally centrally symmetric latent errors.

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. 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 correlations in latent errors of different choices.

Symmetry has been used to obtain point identification in many 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. An example of employing joint symmetry to identify a multiple dependent variable model is the two player entry game of Zhou (2021).[2]

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 derivatives of this function with respect to 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.

[2] We differ from Zhou (2021) in many ways, e.g., we consider 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 in levels of payoffs; and we obtain a new and different estimator based on our identification.

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 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. Many of those methods assume independence between the covariates and error terms, ruling out the possibility of individual heterogeneity such as random coefficients (Ruud (1986), Powell and Ruud (2008), Shi, Shum and Song (2018), and Khan, Ouyang, and Tamer (2019).[3] Some assume exchangeable errors across alternatives (e.g., Manski (1975, 1985), Fox (2007), and Yan (2013)) which impose restrictions on the permitted forms of correlation and heteroskedasticity across alternatives, or assume a limited form of heteroskedasticity (e.g., Lee (1995) and Ahn, Ichimura, Powell, and Ruud (2018)). 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.[4]

This paper is organized as follows. We show identification in Section 2, using the special case of three alternatives. We provide an estimator in Section 3, Monte Carlo simulations in Section 4, and Section 5 concludes. In an online supplementary appendix, we present identification for

[3] Shi, Shum, and Song (2018) and Khan, Ouyang, and Tamer (2019) incorporate unobserved individual heterogeneity by exploiting panel data.

[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 mulitnomial discrete choice models.

the general multinomial choice case, we show root-N consistency and asymptotic normality of our estimator, and we provide proofs for all of our theorems.

2 The Model and Identification

2.1 The Random Utility Framework

To simplify notation and presentation of our results, for the main text of this paper we restrict attention to the case of three choices, with the relative utility of the outside option, denoted $j = 0$, normalized to equal zero. General results for an arbitrary number of multinomial choices, and allowing the outside option to vary, are in the Supplemental Appendix.

For each alternative $j = 0, 1, 2$, let u_j be the difference between the latent utility associated with choice j and the utility of choice zero. Latent utilities u_j are not observed. Assume utility functions

$$u_j = z_j + \boldsymbol{x}_j' \boldsymbol{\theta}^o + \varepsilon_j \text{ for } j = 1, 2, \text{ and } u_0 = 0, \tag{1}$$

where z_j is a continuously distributed covariate with a coefficient normalized to equal one,[5] $\boldsymbol{x}_j = (x_{j1}, x_{j2}, ...x_{jq})'$ is a q vector of other covariates, $\boldsymbol{\theta}^o$ is the q vector of preference parameters of interest, and ε_j is an unobserved random component of utility for alternative j. Errors $\varepsilon_1, \varepsilon_2$ can depend on $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, so in particular we may have random coefficients that are absorbed into ε_1 and ε_2.

Let the dummy variable y_j indicate whether alternative j yields the highest utility among all the alternatives, that is,

$$y_j = \mathbb{I}\left(u_j \geq u_k \quad \forall\, k \neq j\right). \tag{2}$$

Note $y_0 + y_1 + y_2 = 1$. We require that the econometrician observes z_1, z_2, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and y_0. Greater efficiency is possible if y_1 and y_2 are also observed. But identification only requires observing a

[5] In parametric models it is common to normalize the error variance to equal one, but semiparametrically it is often more convenient to normalize a coefficient.

single outcome like y_0, because the choice of any one outcome depends on the utilities of all of the outcomes.

The model can include both individual and alternative specific covariates. E.g., if s is an individual specific covariate (that doesn't vary by choice j), then we could let $\boldsymbol{x}_{11} = \boldsymbol{x}_{22} = s$ and $\boldsymbol{x}_{12} = \boldsymbol{x}_{21} = 0$, so θ_1 and θ_2 would then be the coefficients of s in the utility of choice 1 and 2, respectively. Separate constant terms in u_1 and u_2 can similarly be included in the model. Note z_1 and z_2 are alternative specific covariates.

By equation (2), we have

$$\begin{aligned}\Pr(y_0 = 1 \mid \boldsymbol{z}, \boldsymbol{X}) &= \Pr(u_1 \leq u_0, u_2 \leq u_0 \mid \boldsymbol{z}, \boldsymbol{X}) = \Pr(u_1 \leq 0, u_2 \leq 0 \mid \boldsymbol{z}, \boldsymbol{X}) \qquad (3)\\ &= F_{\varepsilon_1 \varepsilon_2}\left(-z_1 - \boldsymbol{x}_1' \boldsymbol{\theta}^o, -z_2 - \boldsymbol{x}_2' \boldsymbol{\theta}^o \mid \boldsymbol{z}, \boldsymbol{X}\right),\end{aligned}$$

where $\boldsymbol{z} \equiv (z_1, z_2)'$, $\boldsymbol{X} \equiv (\boldsymbol{x}_1, \boldsymbol{x}_2)'$, and $F_{\varepsilon_1 \varepsilon_2}$ is the distribution function of the errors, conditional on covariates. Let sets $\mathcal{S}_{\boldsymbol{z}}$ and $\mathcal{S}_{\boldsymbol{X}}$ denote the supports of the random vector $\boldsymbol{z}$ and random matrix $\boldsymbol{X}$, respectively. Let sets $\mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X})$ and $\mathcal{S}_{\varepsilon_1 \varepsilon_2}(\boldsymbol{X})$ denote the supports of vectors $\boldsymbol{z}$ and $(\varepsilon_1, \varepsilon_2)$ conditional on the values of $\boldsymbol{X}$, respectively.

2.2 Key Conditions For Identification

Here we provide our key assumptions for identification, focusing on the case of three alternatives, i.e., $J = 2$. Identification for the general case $J \geq 2$ is proven in the Supplementary Appendix.

Assumption I.

- **I1**: Conditional on almost every $\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{X}}$, the covariate vector $\boldsymbol{z}$ is independent of the error vector $\boldsymbol{\varepsilon}$, and the conditional distribution function of $\boldsymbol{z}$, $F_{\boldsymbol{z}}(\cdot \mid \boldsymbol{X})$, is absolutely continuous over its support $\mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X})$.[6]

- **I2**: For almost every $\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{X}}$, the conditional distribution function $F_{\varepsilon_1 \varepsilon_2}(t_1, t_2 \mid \boldsymbol{X})$ admits

[6] $\mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X})$ is an open subset of $\mathcal{R}^J$.

an absolutely continuous density function, $f_{\varepsilon_1\varepsilon_2}(t_1, t_2|\, \boldsymbol{X})$, which is centrally symmetric about the origin, i.e.,

$$f_{\varepsilon_1\varepsilon_2}(t_1, t_2|\, \boldsymbol{X}) = f_{\varepsilon_1\varepsilon_2}(-t_1, -t_2|\, \boldsymbol{X}),$$

for any vector $(t_1, t_2) \in \mathcal{S}_{\varepsilon_1\varepsilon_2}(\boldsymbol{X})$ where $\mathcal{S}_{\varepsilon_1\varepsilon_2}(\boldsymbol{X}) \subseteq \mathcal{R}^2$.

- **I3**: The true parameter vector $\boldsymbol{\theta}^o$ is in the parameter space $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a compact set in $\mathcal{R}^q$.

- **I4**: (a) For any constant vector $\boldsymbol{c} = (c_1, \ldots, c_q)' \in \mathcal{R}^q$, $P(\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0}_J) = 1$ if and only if $\boldsymbol{c} = \boldsymbol{0}_q$. (b) The joint density function of the continuous random variables in $\boldsymbol{X}$ is absolutely continuous and positive over its support. For every $\boldsymbol{X}^* \in \mathcal{S}_{\boldsymbol{X}}$, the conditional density function of $\boldsymbol{z}$ given $\boldsymbol{X}$, $f_{\boldsymbol{z}}(\cdot \mid \boldsymbol{X} = \boldsymbol{X}^*)$, is absolutely continuous and positive over its support $\mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X}^*)$. (c) For every $\boldsymbol{X}^* \in \mathcal{S}_{\boldsymbol{X}}$, there exists a subset $\widetilde{\mathcal{S}}_{\boldsymbol{z}}(\boldsymbol{X}^*)$, where $\widetilde{\mathcal{S}}_{\boldsymbol{z}}(\boldsymbol{X}^*) \subseteq int(\mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X}^*))$, with positive measure such that $-\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta} \in int(\mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X}^*))$ for every $\boldsymbol{z}^* \in \widetilde{\mathcal{S}}_{\boldsymbol{z}}(\boldsymbol{X}^*)$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

- **I5:** For every $\boldsymbol{X}^* \in \mathcal{S}_{\boldsymbol{X}}$, $\widetilde{\mathcal{S}}_{\boldsymbol{\varepsilon}}(\boldsymbol{X}^*) \equiv \{-\boldsymbol{z}^* - \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X}^*)\} \cup \{\boldsymbol{z}^* + 2\boldsymbol{X}^*\boldsymbol{\theta} - \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X}^*),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is a subset of the interior of the support $\mathcal{S}_{\boldsymbol{\varepsilon}}(\boldsymbol{X}^*)$. For every $\boldsymbol{X}^* \in \mathcal{S}_{\boldsymbol{X}}$ and any constant vector $\boldsymbol{r} \in \mathcal{R}^J$, $f_{\boldsymbol{\varepsilon}}(\boldsymbol{t} \mid \boldsymbol{X} = \boldsymbol{X}^*) = f_{\boldsymbol{\varepsilon}}(\boldsymbol{r} - \boldsymbol{t} \mid \boldsymbol{X} = \boldsymbol{X}^*)$ for every $\boldsymbol{t} \in \widetilde{\mathcal{S}}_{\boldsymbol{\varepsilon}}(\boldsymbol{X}^*)$ and $\boldsymbol{r} - \boldsymbol{t} \in \widetilde{\mathcal{S}}_{\boldsymbol{\varepsilon}}(\boldsymbol{X}^*)$ if and only if $\boldsymbol{r} = \boldsymbol{0}_J$.

2.3 Identification Strategy

Consider taking the derivatives of both sides of (3) with respect to each elements of $\boldsymbol{z}$, and evaluate the resulting function at the points $(\boldsymbol{z} = \boldsymbol{z}^*, \boldsymbol{X} = \boldsymbol{X}^*)$ and $(\boldsymbol{z} = -\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X} = \boldsymbol{X}^*)$, for some chosen values of $\boldsymbol{z}^*$, $\boldsymbol{X}^*$, and $\boldsymbol{\theta}$. Assume these values are chosen such that $\boldsymbol{X}^* \in \mathcal{S}_{\boldsymbol{X}}$, $\boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X}^*)$, and $-\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta} \in \mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X}^*)$ (which we can do by Assumption I5). By Assumption

I1, this yields the equations

$$\frac{\partial^2 E\left(y_0 \mid \boldsymbol{z}=\boldsymbol{z}^*, \boldsymbol{X}=\boldsymbol{X}^*\right)}{\partial z_1 \partial z_2} = \frac{\partial^2 \Pr\left(y_0=1 \mid \boldsymbol{z}=\boldsymbol{z}^*, \boldsymbol{X}=\boldsymbol{X}^*\right)}{\partial z_1 \partial z_2} \tag{4}$$

$$= f_{\varepsilon_1\varepsilon_2}\left(-z_1^* - \boldsymbol{x}_1^{*\prime}\boldsymbol{\theta}^o, -z_2^* - \boldsymbol{x}_2^{*\prime}\boldsymbol{\theta}^o \mid \boldsymbol{X}=\boldsymbol{X}^*\right) \times (-1)^2,$$

and

$$\frac{\partial^2 E\left(y_0 \mid \boldsymbol{z}=-\boldsymbol{z}^*-2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X}=\boldsymbol{X}^*\right)}{\partial z_1 \partial z_2} = \frac{\partial^2 \Pr\left(y_0=1 \mid \boldsymbol{z}=-\boldsymbol{z}^*-2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X}=\boldsymbol{X}^*\right)}{\partial z_1 \partial z_2} \tag{5}$$

$$= f_{\varepsilon_1\varepsilon_2}\left(z_1^* + 2\boldsymbol{x}_1^{*\prime}\boldsymbol{\theta} - \boldsymbol{x}_1^{*\prime}\boldsymbol{\theta}^o, z_2^* + 2\boldsymbol{x}_2^{*\prime}\boldsymbol{\theta} - \boldsymbol{x}_2^{*\prime}\boldsymbol{\theta}^o \mid \boldsymbol{X}=\boldsymbol{X}^*\right) \times (-1)^2.$$

The left sides of equations (4) and (5) are both identified, and can be estimated as nonparametric regression derivatives, given a value of $\boldsymbol{\theta}$. If $\boldsymbol{\theta}=\boldsymbol{\theta}^o$, then by the symmetry Assumption I2, the right sides of equations (4) and (5) are equal to each other. Define the function $d_0(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*)$ as the difference between the left sides of equations (4) and (5),

$$d_0\left(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*\right) \equiv \frac{\partial^2 E\left(y_0 \mid \boldsymbol{z}=\boldsymbol{z}^*, \boldsymbol{X}=\boldsymbol{X}^*\right)}{\partial z_1 \partial z_2} - \frac{\partial^2 E\left(y_0 \mid \boldsymbol{z}=-\boldsymbol{z}^*-2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X}=\boldsymbol{X}^*\right)}{\partial z_1 \partial z_2}. \tag{6}$$

If $\boldsymbol{\theta}=\boldsymbol{\theta}^o$, then $d_0\left(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*\right)=0$. Given some regularity conditions, we can show that setting the function d_0 equal to zero at a collection of values of $\boldsymbol{z}^*$ and $\boldsymbol{X}^*$ provides enough equations to point identify $\boldsymbol{\theta}^o$. We now formalize this identification strategy.

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

$$\mathcal{D}_0\left(\boldsymbol{\theta}\right) \equiv \left\{\left(\boldsymbol{z}^*,\boldsymbol{X}^*\right) \in int\left(\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})}\right) \middle| \left(-\boldsymbol{z}^*-2\boldsymbol{X}^*\boldsymbol{\theta},\boldsymbol{X}^*\right) \in int\left(\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})}\right), d_0\left(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*\right) \neq 0\right\}. \tag{7}$$

The true parameter vector $\boldsymbol{\theta}^o\in\boldsymbol{\Theta}$ is point identified if $P\left[(\boldsymbol{z},\boldsymbol{X})\in\mathcal{D}_0(\boldsymbol{\theta})\right]=0$ if and only if $\boldsymbol{\theta}=\boldsymbol{\theta}^o$, where $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

At the true parameter vector $\boldsymbol{\theta}^o$, $\mathcal{D}_0(\boldsymbol{\theta}^o)$ is an empty set because $d_0\left(\boldsymbol{\theta}^o;\boldsymbol{z}^*,\boldsymbol{X}^*\right)=0$ for every $(\boldsymbol{z}^*,\boldsymbol{X}^*)$, $(-\boldsymbol{z}^*-2\boldsymbol{X}^*\boldsymbol{\theta}^o,\boldsymbol{X}^*)\in\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})}$. To achieve identification, we require that the set $\mathcal{D}_0(\boldsymbol{\theta})$

have positive probability measure for any $\boldsymbol{\theta}$ in the parameter space other than $\boldsymbol{\theta}^o$. Assumptions I4 and I5 give one set of conditions that suffice.Assumption I4 provides a subset of the support of covariates with positive measure on which the function $d_0(\boldsymbol{\theta}; \boldsymbol{z}^*, \boldsymbol{X}^*)$ can be identified, while Assumption I5 ensures that symmetry points are unique.

Given these assumptions we obtain identification as follows. All proofs are in the Supplementary Appendix.

Theorem 2.1 *If Assumption I hold, then the parameter vector* $\boldsymbol{\theta}^o \in \boldsymbol{\Theta}$ *is point identified by Definition* 2.1*.*

2.3.1 Discussion

Theorem 2.1 used expectations of y_0. Additional identifying information (resulting in more efficient associated estimators) can similarly be obtained from y_1 and y_2. Details are in the supplemental appendix.

The conditional independence between $\boldsymbol{z}$ and $\boldsymbol{\varepsilon}$ in Assumption I1 is known as a distributional exclusion restriction (Powell, 1994, p. 2484). This allows for interpersonal heteroskedasticity on a subset of covariates: Higher moments of $\boldsymbol{\varepsilon}$ can depend (in unknown ways) on $\boldsymbol{X}$, but not $\boldsymbol{z}$. Assumption I2 is our 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 in equation (1).

Assumption I3 assumes a compact parameter space, which is a standard assumption for many nonlinear models, including semiparametric multinomial discrete choice models. Assumption I4(a) is a standard full rank condition, ruling out perfect collinearity among the regressors. Assumption I4(b) requires all continuous covariates to have a positive joint density. Assumption I4(c) guarantees that,given any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the function $d_0(\boldsymbol{\theta}; \boldsymbol{z}^*, \boldsymbol{X}^*)$ can be identified for a positive measure of covariates.

Assumption I5(a) ensures that the error density functions in (4) and (5) are evaluated at interior points of their support. Assumption I5(b) requires that the error density function has a unique (local) symmetry point over a subset of its support, $\widetilde{\mathcal{S}}_{\boldsymbol{\varepsilon}}(\boldsymbol{X}^*)$. This does not rule out densities having flat sections, but it does limit the range of any such flat sections.

2.4 An Alternative Identification Strategy

Existing binary choice estimators that make use of latent error symmetry (e.g. Chen (2000) and Chen, Khan and Tang (2016) are based on the error distribution function rather than on the error density function as in Theorem 2.1. To illustrate, take a simple binary choice model where $y = \mathbb{I}(z + a + v \geq 0)$. If v is a symmetric random variable around zero and $v \perp z$, then

$$\mathbb{E}(y \mid z = c) = \Pr(v \geq -c - a) = \Pr(v \leq c + a) = \mathbb{E}(1 - y \mid z = -c - 2a) \tag{8}$$

The constant a is identified by equating the above two expectations, which only requires estimation of the conditional mean of y and not its derivatives. This immediately extends to identification of covariate coefficients instead of just a constant.

We could have similarly based identification and estimation of our multinomial $\boldsymbol{\theta}$ on the distribution instead of the density of the errors, and thereby only required nonparametric regressions and not their derivatives for estimation. However, unlike the binary choice case, identification and estimation using the distribution instead of the density of the errors becomes complicated and clumsy in the multinomial setting. This is because, in the binary choice case, error symmetry just equates two conditional expectations, corresponding to two error intervals, while for multinomial choice, one must equate error rectangles.

To see the issue, begin again from equation (3). Let $[\boldsymbol{a}, \boldsymbol{b}]$ be a rectangle in the support of $\boldsymbol{\varepsilon}$. Point $\boldsymbol{a} = (a_1, a_2)$ is the lower left vertex of this rectangle and $\boldsymbol{b} = (b_1, b_2)$ is the upper right vertex. By central symmetry, the probability of $\boldsymbol{\varepsilon}$ being in the rectangle $[\boldsymbol{a}, \boldsymbol{b}] = [a_1, b_1] \times [a_2, b_2]$ is the same as the probability of $\boldsymbol{\varepsilon}$ being in the rectangle $[-\boldsymbol{b}, -\boldsymbol{a}] = [-b_1, -a_1] \times [-b_2, -a_2]$. This then implies

$$\int_{[\boldsymbol{a},\boldsymbol{b}]} f_{\varepsilon_1\varepsilon_2}\left(t_1,t_2|\,\boldsymbol{X}\right)d\boldsymbol{t} = \int_{[\boldsymbol{a},\boldsymbol{b}]} f_{\varepsilon_1\varepsilon_2}\left(-t_1,-t_2|\,\boldsymbol{X}\right)d\boldsymbol{t} = \int_{[-\boldsymbol{b},-\boldsymbol{a}]} f_{\varepsilon_1\varepsilon_2}\left(t_1,t_2|\,\boldsymbol{X}\right)d\boldsymbol{t}, \tag{9}$$

where the first equality in (9) holds by Assumption A2 and the second one holds by changing of variables.[7]

The integrals on both sides of equation (9) can be computed using the conditional distribution function of $\boldsymbol{\varepsilon}$, which in turn is obtained from the conditional expectation of y_0. For example, consider the left-hand side integral:

$$\begin{aligned} \textstyle\int_{[\boldsymbol{a},\boldsymbol{b}]} f_{\boldsymbol{\varepsilon}}\left(\boldsymbol{t}\mid\boldsymbol{X}\right)d\boldsymbol{t} &= \Pr\left(\boldsymbol{a}\leq\boldsymbol{\varepsilon}\leq\boldsymbol{b}\mid\boldsymbol{X}\right) \\ &= \Pr\left(a_1\leq\varepsilon_1\leq b_1, a_2\leq\varepsilon_2\leq b_2\mid\boldsymbol{X}\right) \\ &= \Pr\left(\varepsilon_1\leq b_1,\varepsilon_2\leq b_2\mid\boldsymbol{X}\right)-\Pr\left(\varepsilon_1< a_1,\varepsilon_2\leq b_2\mid\boldsymbol{X}\right) \\ &\quad -\Pr\left(\varepsilon_1\leq b_1,\varepsilon_2< a_2\mid\boldsymbol{X}\right)+\Pr\left(\varepsilon_1< a_1,\varepsilon_2< a_2\mid\boldsymbol{X}\right). \end{aligned} \tag{10}$$

The right side of this equation can be rewritten as a function of the conditional expectation of y_0 evaluated at four different points, which in turn means that the multinomial analog to equation (8), obtained from equation (9) requires evaluating the conditional expectation of y_0 evaluated at eight different points functions of $\boldsymbol{z}$, $\boldsymbol{X}$, and θ. This was for our simple multinomial model with 3 choices. The number of required points increases exponentially with the number of choices.

Our density based identification and estimation entails matching points (that is, using $f_{\boldsymbol{\varepsilon}}\left(\boldsymbol{t}\mid\boldsymbol{X}\right) = f_{\boldsymbol{\varepsilon}}\left(-\boldsymbol{t}\mid\boldsymbol{X}\right)$ at data points $\boldsymbol{t}$) 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.[8] In contrast, matching densities rather than distributions at points works for identifying and estimating both binary and multinomial choice, and extends to any number of choices.

[7] Equation (9) also holds when $J>2$, taking $[\boldsymbol{a},\boldsymbol{b}]$ and $[-\boldsymbol{b},-\boldsymbol{a}]$ to be centrally symmetric hyper-rectangles, and it holds for the binary choice model $J=1$, where $[\boldsymbol{a},\boldsymbol{b}]$ and $[-\boldsymbol{b},-\boldsymbol{a}]$ reduce to symmetric intervals about the origin. The identifying binary choice equation given above corresponds to this case with $\boldsymbol{b}=\infty$.

[8] For binary choice we have $J=1$, making $\boldsymbol{\varepsilon}$ and $\boldsymbol{t}$ being scalars, and by symmetry we get $F_{\boldsymbol{\varepsilon}}\left(\boldsymbol{t}\mid\boldsymbol{X}\right) = 1-F_{\boldsymbol{\varepsilon}}\left(-\boldsymbol{t}\mid\boldsymbol{X}\right)$ at data points $\boldsymbol{t}$, but this equality does *not* hold for the multinomial case $J>1$, when $\boldsymbol{\varepsilon}$ and $\boldsymbol{t}$ are vectors rather than scalars.

We prefer to identify and estimate $\boldsymbol{\theta}$ 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 ($\boldsymbol{z} = \boldsymbol{z}^*, \boldsymbol{X} = \boldsymbol{X}^*$) in the data that have matches ($\boldsymbol{z} = -\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X} = \boldsymbol{X}^*$) 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.

3 A Minimum Distance Estimator and its Asymptotic Properties

3.1 Population Objective Functions for Estimation

Given the identification strategy described in Section 2, we develop a minimum distance estimator (hereafter, MD estimator) for $\boldsymbol{\theta}^o \in \boldsymbol{\Theta}$ using the identifying restriction $d_0(\boldsymbol{\theta}^o; \boldsymbol{z}^*, \boldsymbol{X}^*) = 0$, where d_0 is defined by equation (6). Note that the function $d_0(\boldsymbol{\theta}; \boldsymbol{z}^*, \boldsymbol{X}^*)$ is well defined if both points $(\boldsymbol{z}^*, \boldsymbol{X}^*)$ and $(-\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X}^*)$ are in the interior of the support of covariates, $\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})}$. For this reason, we only wish to evaluate the function $d_0(\boldsymbol{\theta}; \boldsymbol{z}^*, \boldsymbol{X}^*)$ at such points.

This can be achieved by multiplying d_0 by a suitable trimming function τ_0. Let $\boldsymbol{X}\overline{\boldsymbol{\theta}}$ and $\boldsymbol{X}\underline{\boldsymbol{\theta}}$ be values in the support of the index $\boldsymbol{X}\boldsymbol{\theta}$ where $\boldsymbol{X}\overline{\boldsymbol{\theta}} << \boldsymbol{X}\underline{\boldsymbol{\theta}}$. We will be trimming values outside the range of these values. Define functions $\tau_0(\cdot)$ and $\varsigma_0(\cdot)$ by

$$\tau_0\left(\boldsymbol{z},\boldsymbol{X};\overline{\boldsymbol{\theta}},\underline{\boldsymbol{\theta}}\right)\equiv\varsigma_0\left(\boldsymbol{z},\boldsymbol{X}\right)\varsigma_0\left(-\boldsymbol{z}-2\boldsymbol{X}\overline{\boldsymbol{\theta}},\boldsymbol{X}\right)\varsigma_0\left(-\boldsymbol{z}-2\boldsymbol{X}\underline{\boldsymbol{\theta}},\boldsymbol{X}\right),\tag{11}$$

and

$$\varsigma_0\left(\boldsymbol{z},\boldsymbol{X}\right)\equiv 1\left(|\boldsymbol{z}|\leq\boldsymbol{c}_1\right)\times 1\left(|\boldsymbol{X}|\leq\boldsymbol{C}_2\right).\tag{12}$$

Here the absolute value of a vector or matrix, $|\cdot|$, is defined as the corresponding vector or matrix of the absolute values of each element, $\boldsymbol{c}_1\in\mathcal{R}^2$ is a vector of trimming constants for the covariate vector $\boldsymbol{z}$, and $\boldsymbol{C}_2\in\mathcal{R}^{2\times q}$ is a matrix of trimming constants for the covariate matrix $\boldsymbol{X}$ such that $(\boldsymbol{c}_1,\boldsymbol{C}_2)$ is in the interior of the support of covariates $\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})}$. Denote $\mathcal{S}_{\boldsymbol{z}}^{Tr}\left(\boldsymbol{X},\overline{\boldsymbol{\theta}},\underline{\boldsymbol{\theta}}\right)$ as the largest set of values $\boldsymbol{z}$ given $\overline{\boldsymbol{\theta}}$, $\underline{\boldsymbol{\theta}}$, and $\boldsymbol{X}$, such that $\mathcal{S}_{\boldsymbol{z}}^{Tr}\left(\boldsymbol{X},\overline{\boldsymbol{\theta}},\underline{\boldsymbol{\theta}}\right)\subset int\left(\mathcal{S}_{\boldsymbol{z}}\left(\boldsymbol{X}\right)\right)$. We assume the following regularity condition on the trimming function $\tau_0(\cdot)$

Assumption TR.

The trimming function $\tau_0\left(\boldsymbol{z},\boldsymbol{X};\overline{\boldsymbol{\theta}},\underline{\boldsymbol{\theta}}\right)$ is strictly positive and bounded on $\mathcal{S}_{\boldsymbol{z}}^{Tr}\left(\boldsymbol{X},\overline{\boldsymbol{\theta}},\underline{\boldsymbol{\theta}}\right)\times int\left(\mathcal{S}_{\boldsymbol{X}}\right)$, and is equal to zero on its complementary set.

The population objective function of our proposed MD estimator is

$$Q_0\left(\boldsymbol{\theta}\right)\equiv\frac{1}{2}\mathbb{E}\left[\tau_0\left(\boldsymbol{z}_n,\boldsymbol{X}_n\right)d_0\left(\boldsymbol{\theta};\boldsymbol{z}_n,\boldsymbol{X}_n\right)\right]^2\tag{13}$$

The sample objective function we define later replaces the expectation in (13) with a sample average and replaces $d_0\left(\boldsymbol{\theta};\boldsymbol{z}_n,\boldsymbol{X}_n\right)$ with an estimator of this function. The following theorem provides population identification through the population objective function.

Theorem 3.1 *If Assumption I and Assumption TR hold, then (i) $Q_0\left(\boldsymbol{\theta}\right)\geq 0$ for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and (ii) $Q_j\left(\boldsymbol{\theta}\right)=0$ if and only if $\boldsymbol{\theta}=\boldsymbol{\theta}^o$.*

3.2 An Estimator

We now provide an estimator for function $d_0(\boldsymbol{\theta}; \boldsymbol{z}_n, \boldsymbol{X}_n)$ in (13) as

$$\hat{d}_{0,-n}(\boldsymbol{\theta}; \boldsymbol{z}_n, \boldsymbol{X}_n) \equiv \hat{\varphi}^{(2)}_{o,-n}(\boldsymbol{z}_n, \boldsymbol{X}_n) - \hat{\varphi}^{(2)}_{cs,-n}(\boldsymbol{z}_n, \boldsymbol{X}_n, \boldsymbol{\theta}). \tag{14}$$

where $\hat{\varphi}^{(2)}_{o,-n}(\boldsymbol{z}_n, \boldsymbol{X}_n)$ and $\hat{\varphi}^{(2)}_{cs,-n}(\boldsymbol{z}_n, \boldsymbol{X}_n, \boldsymbol{\theta})$ are leave-one-out, Nadaraya-Watson nonparametric regression kernel estimators for the derivatives on the right hand side of equation (6) (see the supplemental appendix for details). By replacing the expectation in $Q_0(\boldsymbol{\theta})$ with its sample mean and replacing the function $d_0(\boldsymbol{\theta}; \boldsymbol{z}_n, \boldsymbol{X}_n)$ with the estimator $\hat{d}_{0,-n}(\boldsymbol{\theta}; \boldsymbol{z}_n, \boldsymbol{X}_n)$, we define the minimum distance (MD) estimator

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} Q_{N0}(\boldsymbol{\theta}),$$

$$\text{where} \qquad Q_{N0}(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{n=1}^{N}\left[\tau_0(\boldsymbol{z}_n, \boldsymbol{X}_n)\,\hat{d}_{0,-n}(\boldsymbol{\theta}; \boldsymbol{z}_n, \boldsymbol{X}_n)\right]^2.$$

We denote the gradient of the objective function as $\boldsymbol{q}_{N0}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} Q_{N0}(\boldsymbol{\theta})$ and the Hessian matrix of the objective function as $\boldsymbol{H}_{N0}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} Q_{N0}(\boldsymbol{\theta})$. The smoothness of the objective function suggests the first-order condition (FOC): $\boldsymbol{q}_{N0}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{0}_q$. Applying the standard first-order Taylor expansion to $\boldsymbol{q}_{N0}\left(\hat{\boldsymbol{\theta}}\right)$ around the true parameter vector $\boldsymbol{\theta}^o$ yields

$$\boldsymbol{q}_{N0}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{q}_{N0}(\boldsymbol{\theta}^o) + \boldsymbol{H}_{N0}\left(\tilde{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o\right),$$

where $\tilde{\boldsymbol{\theta}}$ is a vector between the MD estimator $\hat{\boldsymbol{\theta}}$ and the true parameter vector $\boldsymbol{\theta}^o$. The influence function is then given by

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o = -\left[\boldsymbol{H}_{N0}\left(\tilde{\boldsymbol{\theta}}\right)\right]^{-1}\boldsymbol{q}_{N0}(\boldsymbol{\theta}^o). \tag{15}$$

In the Supplemental Appendix, we list regularity assumptions and use them to show that $\boldsymbol{H}_{N0}\left(\tilde{\boldsymbol{\theta}}\right) \to_p$

$\boldsymbol{H}_0(\boldsymbol{\theta}^o)$, where

$$\boldsymbol{H}_0(\boldsymbol{\theta}^o) = \mathbb{E}\left\{\tau_0^2(\boldsymbol{z}_n, \boldsymbol{X}_n) \nabla_{\boldsymbol{\theta}} d_0(\boldsymbol{\theta}^o; \boldsymbol{z}_n, \boldsymbol{X}_n) [\nabla_{\boldsymbol{\theta}} d_0(\boldsymbol{\theta}^o; \boldsymbol{z}_n, \boldsymbol{X}_n)]'\right\}, \tag{16}$$

and we show $\sqrt{N}\boldsymbol{q}_{N0}(\boldsymbol{\theta}^o) \to_d N(\mathbf{0}_q, \boldsymbol{\Omega}_0)$, where $\boldsymbol{\Omega}_0$ is the probability limit of the variance-covariance matrix of $\boldsymbol{q}_{N0}(\boldsymbol{\theta}^o)$. More precisely, we prove the following theorems.

Theorem 3.2 *Under Assumption I and TR as well as certain regularity conditions, the MD estimator $\hat{\boldsymbol{\theta}}$ converges to the true parameter vector $\boldsymbol{\theta}^o \in \boldsymbol{\Theta}$ in probability.*

Theorem 3.3 *Under Assumption I and TR as well as certain regularity conditions, we have*

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

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o\right) = -N^{-1/2}\sum_{n=1}^{N} \boldsymbol{H}_0^{-1}\boldsymbol{t}_n + o_p(1),$$

where $\boldsymbol{t}_{n0} \equiv (\upsilon_{n0,o} - \upsilon_{n0,cs})\, \partial^2\left[\tau_0^2(\boldsymbol{z}_n, \boldsymbol{X}_n) \nabla_{\boldsymbol{\theta}} d_0(\boldsymbol{\theta}^o; \boldsymbol{z}_n, \boldsymbol{X}_n)\right]/\partial z_1 \partial z_2$ with scalars $\upsilon_{n0,o} \equiv y_{n0} - \varphi_{0,o}(\boldsymbol{z}_n, \boldsymbol{X}_n)$ and $\upsilon_{n0,cs} \equiv y_{nj} - \varphi_{0,cs}(\boldsymbol{z}_n, \boldsymbol{X}_n, \boldsymbol{\theta}^o)$.

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

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o\right) \to_d N\left(\mathbf{0}_q, \boldsymbol{H}_0^{-1}\boldsymbol{\Omega}_0\boldsymbol{H}_0^{-1}\right)$$

where the matrix $\boldsymbol{\Omega}_0 \equiv \mathbb{E}(\boldsymbol{t}_{n0}\boldsymbol{t}_{n0}')$ and $\boldsymbol{H}_0$ is defined above.

Note that this estimator only makes use of observations of y_0, and not the other choices y_j, and so can be applied if the researcher only observes whether the outside option ($j = 0$) is chosen or not. In the Supplemental Appendix, we extend this estimator to make use of comparable identifying information in all of the other choices (thereby increasing efficiency). This essentially just consists of renormalizing each choice to be the base choice, constructing the above MD estimator (call it $Q_{Nj}(\boldsymbol{\theta})$ for choice j), and then minimizing the sum of the resulting MD objective functions

$Q_{Nj}(\boldsymbol{\theta})$ over each choice. We also extend all our results to multnomial choice with an arbitrary number of choices, instead of just three as above.

4 Monte Carlo Experiments

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

$$u_{nj} = z_{nj} + x_{nj}\theta_n + \varepsilon_{nj} \text{ for } n = 1, 2, ..., N \text{ and } j = 0, 1, 2. \tag{17}$$

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 DGPs 1 and 2, $\theta_n = \theta^o = 0.2$, a constant, while in DGPs 3 and 4 $\theta_n = \theta^o + \delta_n$ where δ_n is a random variable. In DGP 3, the random coefficient θ_n is independent of all the covariates, while in DGP 4, θ_n depends on individuals' characteristics. We focus on estimation of θ^o, which (by our assumed symmetry) equals both the median and the mean of the second coefficient θ_n under all DGPs.

The researcher observes attributes z_{nj} and x_{nj}. We consider both the MD estimator that only uses y_0, and the estimator that uses y_0, y_1, and y_2. We compare these MD estimators to flexible multinomial probit (MNP) with no constraint on the multivariate normal variance matrix. This MNP specification requires estimating θ^o and three parameters of the error vector variance-covariance matrix.

Details of the Monte Carlo design are given in the Supplementary Appendix. However, note that under DGP 1, MNP is correctly specified and the MNP estimator is efficient, so comparisons with MNP show the efficiency loss that comes from using our estimator. Under DGP 2, MNP is misspecified because the errors and covariates are not independent. Under DGP 3, MNP is misspecified but random coefficients MNP is correctly specified. Under DGP 4, both MNP and random coefficient MNP are misspecified, because the random coefficient distribution depends on

Table 1: Monte Carlo Results of estimating θ^o (True Parameter $\theta^o = 0.2$)

		MNP		MD (y_0)		MD (y_0, y_1, y_2)	
DGP	N	Bias	RMSE	Bias	RMSE	Bias	RMSE
1	1000	-0.0012	0.0435	0.0216	0.2368	-0.0017	0.1337
	2000	-0.0010	0.0307	0.0055	0.1355	-0.0078	0.0788
2	1000	0.5656	0.5833	0.1047	0.3521	-0.0392	0.3048
	2000	0.5627	0.5714	0.0543	0.2308	-0.0289	0.1747
3	1000	-0.0013	0.0454	0.0317	0.2220	0.0015	0.1417
	2000	-0.0017	0.0319	0.0158	0.1301	-0.0051	0.0812
4	1000	-0.7512	0.7718	-0.0054	0.3765	-0.0748	0.3550
	2000	-0.7481	0.7585	0.0180	0.2616	-0.0343	0.2149

covariates. Under all four DGP's our MD estimator remains consistent.

Table 1 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 y_0, while the third uses observations of all choices y_0, y_1, and 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 efficient MNP estimator. Under DGP 2, the bias of the misspecified 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, in DGP 4 where the random component is heterogeneous, 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 relatively small.[9] In all the DGPs, in terms of RMSE, the MD estimator using y_0, y_1, and y_2 performs better than

[9] 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.

the MD estimator that only uses y_0.

Our Monte Carlo experiments study our estimator, but also provide evidence regarding the reliability of MNP, 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 Pertrin, 2004). In multinomial discrete choice, both unobserved choice attributes and individual heterogeneity add complexity to the error structure. Our results show that ignoring either one may result in MNP being severely biased.

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. 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).

References

[1] Ahn, H., Ichimura, H., Powell, J.L., and Ruud, P. (2018): "Simple Estimators for Invertible Index Models," *Journal of Business and Economic Statistics*, 36, 1-10.

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

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

[4] Berry, S., Levinsohn, J., and Pakes, A. (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112(1), 68-105.

[5] Blundell, R. and Powel, J.L. (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655–679.

[6] Chen, S. (2000): "Efficient Estimation of Binary Choice Models under Symmetry," *Journal of Econometrics*, 96, 183-199.

[7] Chen, S. and Zhou, Y. (2010): "Semiparametric and Nonparametric Estimation of Sample Selection Models under Symmetry," *Journal of Econometrics*, 157, 143-150.

[8] Chen, X., Hu,Y., and Lewbel, A. (2008): " Nonparametric Identification of Regression Models Containing a Misclassified Dichotomous Regressor Without Instruments," *Economics Letters*, 100, 381-384.

[9] Chen, S., Khan, S., and Tang, X. (2016): "Informational Content of Special Regressors in Heteroskedastic Binary Response Models," *Journal of Econometrics*, 193, 162-182.

[10] Delaigle, A. and Hall, P. (2016): "Methodology for Non-parametric Deconvolution When the Error Distribution is Unknown," *Journal of the Royal Statistical Society*, Series B, 78, 231–252.

[11] Dong, Y. and Lewbel, A. (2011): "Nonparametric Identification of a Binary Random Factor in Cross Section Data," *Journal of Econometrics*, 163, 163-171.

[12] Fox, J.T. (2007): "Semiparametric Estimation of Multinomial Discrete-choice Models Using a Subset of Choices," *RAND Journal of Economics*, 38, 1002-1019.

[13] Fox, J.T. and Gandhi, A. (2016): "Nonparametric Identification and Estimation of Random Coefficients in Multinomial Choice Models," *RAND Journal of Economics*, 47, 118-139.

[14] Goolsbee, A. and Petrin. A (2004): "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 72, 351-381.

[15] Hausman, J.A. and Wise, D.A. (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46, 403-426.

[16] Heckman, J. (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture" *Journal of Political Economy*, 109, 673-748.

[17] Honoré, B. E., Kyriazidou, E., and Udry, C. (1997): "Estimation of Type 3 Tobit Models using Symmetric Trimming and Pairwise Comparisons," *Journal of Econometrics*, 76, 107-128.

[18] Khan, S., Ouyang, F., and Tamer, E. (2019): " Inference in Semiparametric Multinomial Response Models," Working Paper.

[19] Kumbhakar, S.C. and Lovell, C.A.K. (2000): Stochastic Frontier Analysis. Cambridge University Press, Cambridge.

[20] Lee, L. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381-428.

[21] Lewbel, A. (1997): "Constructing Instruments for Regressions With Measurement Error When no Additional Data are Available, with An Application to Patents and R&D," *Econometrica*, 65, 1201-1213.

[22] Lewbel, A. (2000): "Semiparametric Qualitative Response Model Estimation with Unknown Heteroskedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145-177.

[23] Manski, C.F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205-228.

[24] Manski CF. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313-333.

[25] Manski, C.F. (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, 729-738.

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

[27] McFadden, D. and Train, K. (2000): "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15, 447-470.

[28] Powell, J.L. (1986a): "Censored Regression Qantiles," *Journal of Econometrics*, 32, 143-155.

[29] Powell, J.L. (1986b): "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54, 1435-1460.

[30] Powell, J.L. (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics*, Vol. 4. ed. by R.F.Engle and D.L.McFadden. Elsevier Science B. V. 2443-2521.

[31] Powell, J.L. and Ruud, P.A. (2008): "Simple Estimators for Semiparametric Multinomial Choice Models," Working Paper.

[32] Ruud, P.A. (1986): "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics*, 32, 157-187. Cambridge University Press.

[33] Serfling, R. (2006): "Multivariate Symmetry and Asymmetry," Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc.

[34] Shi, X., Shum, M., and Song, W. (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity," *Econometrica*, 86, 737-761.

[35] Train, K. (2009): Discrete Choice Methods with Simulation.

[36] Yan, J. (2013): "A Smoothed Maximum Score Estimator for Multinomial Discrete Choice Models," Working Paper.

[37] Yan, J. and Yoo, H. (2019): "Semiparametric Estimation of the Random Utility Model with Rank-ordered Choice Data," *Journal of Econometrics*, 211, 414-438.

[38] Zhou, Y. (2021): "Identification and Estimation of of Entry Games under the Symmetry of Unobservables," Working Paper, NYU, Shanghai.

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

Arthur Lewbel[†] Jin Yan[‡] Yu Zhou[§]

Original February 2019, revised December 2021

S.A The General Model and Identification

S.A.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 $\tilde{z}_j \in \mathcal{R}$ and $\tilde{\boldsymbol{x}}_j \in \mathcal{R}^q$ 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 $\tilde{u}_j$ from choosing alternative $j \in \mathbb{J}$ is assumed to be given by:

$$\tilde{u}_j = \tilde{z}_j \gamma^o + \tilde{\boldsymbol{x}}_j' \boldsymbol{\theta}^o + \tilde{\varepsilon}_j \quad \forall\, j \in \mathbb{J}, \tag{S.A.1}$$

where $\gamma^o \in \mathcal{R}$ and $\boldsymbol{\theta}^o \in \mathcal{R}^q$ are the preference parameters of interest, and $\tilde{\varepsilon}_j \in \mathcal{R}$ is the unobserved random component of utility for alternative j. The utility index $\tilde{z}_j \gamma^o + \tilde{\boldsymbol{x}}_j' \boldsymbol{\theta}^o$ is often called

*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.

[§]Economics, New York University Shanghai; Shanghai. Email: amanda.yu.zhou@nyu.edu

systematic (or deterministic) utility, as opposed to the error term, $\tilde{\varepsilon}_j$, 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,

$$y_j = \mathbb{I}\left(\tilde{u}_j \geq \tilde{u}_k \quad \forall\, k \in \mathbb{J} \setminus \{j\}\right). \tag{S.A.2}$$

The choice of the individual is denoted $\boldsymbol{y} \equiv (y_0, y_1, \ldots, y_J)$, where $\sum_{j=0}^{J} y_j = 1$. The econometrician observes the covariates $\tilde{z}_j$ and $\tilde{\boldsymbol{x}}_j$ 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 $\boldsymbol{y}$ as well. The latent utility vector $\tilde{\boldsymbol{u}} \equiv (\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_J)$ is not observed.[1]

Only differences in utilities matter in making choice decisions. Without loss of generality, we set alternative $0 \in \mathbb{J}$ 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 $\boldsymbol{u} \equiv (u_1, \ldots, u_J)' \in \mathcal{R}^J$, where $u_j = \tilde{u}_j - \tilde{u}_0$. For each alternative $j = 1, \ldots, J$, we have the utility function

$$u_j = z_j + \boldsymbol{x}_j' \boldsymbol{\theta}^o + \varepsilon_j, \tag{S.A.3}$$

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

$$\boldsymbol{u} = \boldsymbol{z} + \boldsymbol{X}\boldsymbol{\theta}^o + \boldsymbol{\varepsilon}, \tag{S.A.4}$$

where $\boldsymbol{z} \equiv (z_1, \ldots, z_J)' \in \mathcal{R}^J$, $\boldsymbol{X} \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)' \in \mathcal{R}^{J \times q}$, and $\boldsymbol{\varepsilon} \equiv (\varepsilon_1, \ldots, \varepsilon_J)' \in \mathcal{R}^J$. We use the compact expressions of utility functions defined in (S.A.3) and (S.A.4) 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 $\tilde{\boldsymbol{x}}_0$ are zero, making z_j and $\boldsymbol{x}_j$ equal $\tilde{z}_j$ and $\tilde{\boldsymbol{x}}_j$. Regardless,

[1] 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.

we only require that differences $\boldsymbol{z}$ and $\boldsymbol{X}$ be observed, and regularity conditions (e.g., continuity of $\boldsymbol{z}$) are only be imposed on z_j and $\boldsymbol{x}_j$, not on $\tilde{z}_j$ and $\tilde{\boldsymbol{x}}_j$. In addition to these covariates, our identification only requires that y_0 be observed, not the entire vector of outcomes $\boldsymbol{y}$. This is possible because $\boldsymbol{z}$ provides information about the other outcomes. Nevertheless, the associated estimators will be more efficient by observing and making use of more the elements of $\boldsymbol{y}$, since each additional outcome y_j one observes provides additional overidentifying information.

Assumption I1 immediately implies

$$\begin{aligned}\Pr(y_0 = 1 \mid \boldsymbol{z}, \boldsymbol{X}) &= F_{\varepsilon_1\varepsilon_2\cdots\varepsilon_J}\left(-z_1 - \boldsymbol{x}_1'\boldsymbol{\theta}^o, -z_2 - \boldsymbol{x}_2'\boldsymbol{\theta}^o, ..., -z_J - \boldsymbol{x}_J'\boldsymbol{\theta}^o \mid \boldsymbol{z}, \boldsymbol{X}\right) \\ &= F_{\varepsilon_1\varepsilon_2\cdots\varepsilon_J}\left(-z_1 - \boldsymbol{x}_1'\boldsymbol{\theta}^o, -z_2 - \boldsymbol{x}_2'\boldsymbol{\theta}^o, ..., -z_J - \boldsymbol{x}_J'\boldsymbol{\theta}^o | \boldsymbol{X}\right).\end{aligned} \tag{S.A.5}$$

where the second equality holds by the conditional independence between $\boldsymbol{z}$ and $\boldsymbol{\varepsilon}$ in Assumption I1. In addition, Assumption I1 yields the equations

$$\begin{aligned}\frac{\partial^J E(y_0 \mid \boldsymbol{z} = \boldsymbol{z}^*, \boldsymbol{X} = \boldsymbol{X}^*)}{\partial z_1 \ldots \partial z_J} &= \frac{\partial^J \Pr(y_0 = 1 \mid \boldsymbol{z} = \boldsymbol{z}^*, \boldsymbol{X} = \boldsymbol{X}^*)}{\partial z_1 \ldots \partial z_J} \\ &= f_{\boldsymbol{\varepsilon}}(-\boldsymbol{z}^* - \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{X} = \boldsymbol{X}^*) \times (-1)^J,\end{aligned} \tag{S.A.6}$$

and

$$\begin{aligned}\frac{\partial^J E(y_0 \mid \boldsymbol{z} = -\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X} = \boldsymbol{X}^*)}{\partial z_1 \ldots \partial z_J} &= \frac{\partial^J \Pr(y_0 = 1 \mid \boldsymbol{z} = -\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X} = \boldsymbol{X}^*)}{\partial z_1 \ldots \partial z_J} \\ &= f_{\boldsymbol{\varepsilon}}(\boldsymbol{z}^* + 2\boldsymbol{X}^*\boldsymbol{\theta} - \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{X} = \boldsymbol{X}^*) \times (-1)^J.\end{aligned} \tag{S.A.7}$$

Observe that the left sides of equations (S.A.6) and (S.A.7) are both identified, and can be readily estimated as nonparametric regression derivatives, given $\boldsymbol{\theta}$. It then follows from Assumption I2 that if $\boldsymbol{\theta} = \boldsymbol{\theta}^o$, then the right sides of equations (S.A.6) and (S.A.7) are equal to each other. Therefore, define function $d_0(\boldsymbol{\theta}; \boldsymbol{z}^*, \boldsymbol{X}^*)$ as the difference between the left sides of equations (S.A.6) and (S.A.7),

$$d_0(\boldsymbol{\theta}; \boldsymbol{z}^*, \boldsymbol{X}^*) \equiv \frac{\partial^J E(y_0 \mid \boldsymbol{z} = \boldsymbol{z}^*, \boldsymbol{X} = \boldsymbol{X}^*)}{\partial z_1 \ldots \partial z_J} - \frac{\partial^J E(y_0 \mid \boldsymbol{z} = -\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X} = \boldsymbol{X}^*)}{\partial z_1 \ldots \partial z_J}. \tag{S.A.8}$$

Based on Assumptions I1 and I2, we have that if $\boldsymbol{\theta} = \boldsymbol{\theta}^o$, then $d_0(\boldsymbol{\theta}; \boldsymbol{z}^*, \boldsymbol{X}^*) = 0$. Given some regularity conditions, setting the function d_0 equal to zero at a collection of values of $\boldsymbol{z}^*$ and $\boldsymbol{X}^*$ provides enough equations to point identify $\boldsymbol{\theta}^o$. The proof of Theorem S.A.1 is provided in Section S.C.

Theorem S.A.1 *If Assumptions* I *hold, then the parameter vector* $\boldsymbol{\theta}^o \in \boldsymbol{\Theta}$ *is point identified by Definition* 1.

S.A.2 Identification Using Multiple Choices

In Section A.1, we identified the parameter vector $\boldsymbol{\theta}^o$ 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 $\boldsymbol{X}^{(j)}$ as the matrix that consists of differenced covariate vectors $\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}_j$ for all $k \in \mathbb{J}$ and $k \neq j$. For example, when $1 < j < J$, $\boldsymbol{X}^{(j)} \equiv (\tilde{\boldsymbol{x}}_0 - \tilde{\boldsymbol{x}}_j, \ldots, \tilde{\boldsymbol{x}}_{j-1} - \tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_{j+1} - \tilde{\boldsymbol{x}}_j, \ldots, \tilde{\boldsymbol{x}}_J - \tilde{\boldsymbol{x}}_j)' \in \mathcal{R}^{J \times q}$. By this notation, we have $\boldsymbol{X}^{(0)} \equiv (\tilde{\boldsymbol{x}}_1 - \tilde{\boldsymbol{x}}_0, \ldots, \tilde{\boldsymbol{x}}_J - \tilde{\boldsymbol{x}}_0) = \boldsymbol{X}$. In the same fashion, define $\boldsymbol{z}^{(j)} \in \mathcal{R}^J$ as the vector of differenced covariates $\tilde{z}_k - \tilde{z}_j$ for all $k \neq j$ and $\boldsymbol{\varepsilon}^{(j)} \in \mathcal{R}^J$ as the vector of differenced error terms $\tilde{\varepsilon}_k - \tilde{\varepsilon}_j$ for all $k \neq j$, respectively. By this definition, we have $\boldsymbol{z}^{(0)} = \boldsymbol{z}$ and $\boldsymbol{\varepsilon}^{(0)} = \boldsymbol{\varepsilon}$. Define $\boldsymbol{u}^{(j)} \in \mathcal{R}^J$ as the vector of differenced utilities $\tilde{u}_k - \tilde{u}_j$ for all $k \neq j$. Differenced utility vectors are then given by

$$\boldsymbol{u}^{(j)} = \boldsymbol{z}^{(j)} + \boldsymbol{X}^{(j)}\boldsymbol{\theta}^o + \boldsymbol{\varepsilon}^{(j)}. \tag{S.A.9}$$

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

$$\begin{aligned} P(y_j = 1 \mid \boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}) &= P(\tilde{u}_k - \tilde{u}_j \leq 0 \quad \forall k \in \mathbb{J} \setminus \{j\} \mid \boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}) \\ &= P(\boldsymbol{u}^{(j)} \leq \boldsymbol{0}_J \mid \boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}) \\ &= P(\boldsymbol{\varepsilon}^{(j)} \leq -\boldsymbol{z}^{(j)} - \boldsymbol{X}^{(j)}\boldsymbol{\theta}^o \mid \boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}) \\ &\equiv F_{\boldsymbol{\varepsilon}^{(j)}}(-\boldsymbol{z}^{(j)} - \boldsymbol{X}^{(j)}\boldsymbol{\theta}^o \mid \boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}), \end{aligned} \tag{S.A.10}$$

where the right-hand side of (S.A.10) is the distribution function of the error vector $\boldsymbol{\varepsilon}^{(j)}$ evaluated at the point $-\boldsymbol{z}^{(j)} - \boldsymbol{X}^{(j)}\boldsymbol{\theta}^o$, conditional on covariates $(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)})$. Let sets $\mathcal{S}_{\boldsymbol{z}^{(j)}} \subseteq \mathcal{R}^J$ and $\mathcal{S}_{\boldsymbol{X}^{(j)}} \subseteq \mathcal{R}^{J \times q}$ denote the supports of the random vector $\boldsymbol{z}^{(j)}$ and random matrix $\boldsymbol{X}^{(j)}$, respectively. Let sets $\mathcal{S}_{\boldsymbol{z}^{(j)}}(\boldsymbol{X}^{(j)})$ and $\mathcal{S}_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{X}^{(j)})$ denote the supports of vectors $\boldsymbol{z}^{(j)}$ and $\boldsymbol{\varepsilon}^{(j)}$ conditional on the values of $\boldsymbol{X}^{(j)}$, respectively.

Proposition S.A.1 *If Assumption I1 holds, then for every $j \in \mathbb{J}$ and conditional on almost every $\boldsymbol{X}^{(j)} \in \mathcal{S}_{\boldsymbol{X}^{(j)}}$, covariate vector $\boldsymbol{z}^{(j)}$ is independent of the error vector $\boldsymbol{\varepsilon}^{(j)}$, i.e., $\left(\boldsymbol{z}^{(j)} \perp \boldsymbol{\varepsilon}^{(j)}\right) \mid \boldsymbol{X}^{(j)}$. The distribution function $F_{\boldsymbol{z}^{(j)}}(\cdot \mid \boldsymbol{X}^{(j)})$, is absolutely continuous over its support $\mathcal{S}_{\boldsymbol{z}^{(j)}}(\boldsymbol{X}^{(j)})$.*

Proposition S.A.1 is an immediate result of the fact that there is a one-to-one correspondence between $\boldsymbol{X}^{(j)}$ and $\boldsymbol{X}$, $\boldsymbol{z}^{(j)}$ and $\boldsymbol{z}$, and $\boldsymbol{\varepsilon}^{(j)}$ and $\boldsymbol{\varepsilon}$, respectively. Hence we have

$$\begin{aligned} E(y_j \mid \boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}) &= P(y_j = 1 \mid \boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}) = P(\boldsymbol{\varepsilon}^{(j)} \leq -\boldsymbol{z}^{(j)} - \boldsymbol{X}^{(j)}\boldsymbol{\theta}^o \mid \boldsymbol{X}^{(j)}) \\ &\equiv F_{\boldsymbol{\varepsilon}^{(j)}}(-\boldsymbol{z}^{(j)} - \boldsymbol{X}^{(j)}\boldsymbol{\theta}^o \mid \boldsymbol{X}^{(j)}), \end{aligned} \tag{S.A.11}$$

where the first equality in (S.A.11) holds because y_j is a dummy variable, and the second one holds by (S.A.10) and Proposition S.A.1. For every $\boldsymbol{t} \in \mathcal{S}_{\boldsymbol{\varepsilon}}(\boldsymbol{X})$, let $\boldsymbol{t}^{(j)}$ denote the vector such that $\boldsymbol{\varepsilon}^{(j)} = \boldsymbol{t}^{(j)}$ when $\boldsymbol{\varepsilon} = \boldsymbol{t}$. Given any $\boldsymbol{t}^{(j)} \in \mathcal{S}_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{X}^{(j)})$ we can calculate the conditional mean of y_j on the left-hand side of (S.A.11) at $\boldsymbol{z}^{(j)} = -\boldsymbol{t}^{(j)} - \boldsymbol{X}^{(j)}\boldsymbol{\theta}^o$ as

$$E(y_j \mid \boldsymbol{z}^{(j)} = -\boldsymbol{t}^{(j)} - \boldsymbol{X}^{(j)}\boldsymbol{\theta}^o, \boldsymbol{X}^{(j)}) = P(\boldsymbol{\varepsilon}^{(j)} \leq \boldsymbol{t}^{(j)} \mid \boldsymbol{X}^{(j)}) \equiv F_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{t}^{(j)} \mid \boldsymbol{X}^{(j)}). \tag{S.A.12}$$

Proposition S.A.2 *If Assumption I2 holds, then for every $j \in \mathbb{J}$ and almost every $\boldsymbol{X}^{(j)} \in \mathcal{S}_{\boldsymbol{X}^{(j)}}$, the conditional distribution function $F_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{t}^{(j)} \mid \boldsymbol{X}^{(j)})$ admits an absolutely continuous density function, $f_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{t}^{(j)} \mid \boldsymbol{X}^{(j)})$, which is centrally symmetric about the origin, i.e.,*

$$f_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{t}^{(j)} \mid \boldsymbol{X}^{(j)}) = f_{\boldsymbol{\varepsilon}^{(j)}}(-\boldsymbol{t}^{(j)} \mid \boldsymbol{X}^{(j)}), \tag{S.A.13}$$

for any vector $\boldsymbol{t}^{(j)} \in \mathcal{S}_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{X}^{(j)})$ where $\mathcal{S}_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{X}^{(j)}) \subseteq \mathcal{R}^J$.

To show Proposition S.A.2, observe that for any $\boldsymbol{t} \in \mathcal{S}_{\boldsymbol{\varepsilon}}(\boldsymbol{X})$, we have $\boldsymbol{\varepsilon} = \boldsymbol{t}$ if and only if $\boldsymbol{\varepsilon}^{(j)} = \boldsymbol{t}^{(j)}$ by the one-to-one correspondence between $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^{(j)}$. Therefore,

$$f_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{t}^{(j)} \mid \boldsymbol{X}^{(j)}) = f_{\boldsymbol{\varepsilon}}(\boldsymbol{t} \mid \boldsymbol{X}) = f_{\boldsymbol{\varepsilon}}(-\boldsymbol{t} \mid \boldsymbol{X}) = f_{\boldsymbol{\varepsilon}^{(j)}}(-\boldsymbol{t}^{(j)} \mid \boldsymbol{X}^{(j)}), \tag{S.A.14}$$

where the second equality in (S.A.14) holds by Assumption I2.

Now the remaining derivations mimic that of Theorem S.A.1. Taking the J^{th} order derivatives of both sides of (S.A.11) with respect to each element of $\boldsymbol{z}^{(j)}$ and evaluating them at $(\boldsymbol{z}^{(j)} = \boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)} = \boldsymbol{X}^{(j)*})$ and $(\boldsymbol{z}^{(j)} = -\boldsymbol{z}^{(j)*} - 2\boldsymbol{X}^{(j)*}\boldsymbol{\theta}, \boldsymbol{X}^{(j)} = \boldsymbol{X}^{(j)*})$, respectively, we obtain the equations

$$\begin{aligned} &\partial^J E(y_j \mid \boldsymbol{z}^{(j)} = \boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)} = \boldsymbol{X}^{(j)*})/\partial z_1^{(j)} \ldots \partial z_J^{(j)} \\ &= f_{\boldsymbol{\varepsilon}^{(j)}}(-\boldsymbol{z}^{(j)*} - \boldsymbol{X}^{(j)*}\boldsymbol{\theta}^o \mid \boldsymbol{X}^{(j)} = \boldsymbol{X}^{(j)*}) \times (-1)^J \end{aligned} \tag{S.A.15}$$

and

$$\begin{aligned} &\partial^J E(y_j \mid \boldsymbol{z}^{(j)} = -\boldsymbol{z}^{(j)*} - 2\boldsymbol{X}^{(j)*}\boldsymbol{\theta}, \boldsymbol{X}^{(j)} = \boldsymbol{X}^{(j)*})/\partial z_1^{(j)} \ldots \partial z_J^{(j)} \\ &= f_{\boldsymbol{\varepsilon}^{(j)}}(\boldsymbol{z}^{(j)*} + 2\boldsymbol{X}^{(j)*}\boldsymbol{\theta} - \boldsymbol{X}^{(j)*}\boldsymbol{\theta}^o \mid \boldsymbol{X}^{(j)} = \boldsymbol{X}^{(j)*}) \times (-1)^J. \end{aligned} \tag{S.A.16}$$

By symmetry, if $\boldsymbol{\theta} = \boldsymbol{\theta}^o$ then the two error densities on the right-hand sides of (S.A.15) and (S.A.16) are identical, which implies equality of their left-hand sides. So for any vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $(\boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)*})$, $(-\boldsymbol{z}^{(j)*} - 2\boldsymbol{X}^{(j)*}\boldsymbol{\theta}, \boldsymbol{X}^{(j)*}) \in \mathcal{S}_{(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)})}$, define $d_j(\boldsymbol{\theta}; \boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)*})$ as the difference of the left-hand sides of (S.A.15) and (S.A.16), that is,

$$d_j(\boldsymbol{\theta}; \boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)*}) \equiv \partial^J E(y_j \mid \boldsymbol{z}^{(j)} = \boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)} = \boldsymbol{X}^{(j)*})/\partial z_1^{(j)} \ldots \partial z_J^{(j)} \tag{S.A.17}$$

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

which always equals zero when $\boldsymbol{\theta} = \boldsymbol{\theta}^o$ and may be non-zero when $\boldsymbol{\theta} \neq \boldsymbol{\theta}^o$.

Then, analogous to Definition 1, define

$$\mathcal{D}_j(\boldsymbol{\theta}) \equiv \Big\{ (\boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)*}) \in int\left(\mathcal{S}_{(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)})}\right) \\ \Big| \, (-\boldsymbol{z}^{(j)*} - 2\boldsymbol{X}^{(j)*}\boldsymbol{\theta}, \boldsymbol{X}^{(j)*}) \in int\left(\mathcal{S}_{(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)})}\right), \; d_j(\boldsymbol{\theta}; \boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)*}) \neq 0 \Big\}. \tag{S.A.18}$$

Recall that there is a one-to-one correspondence, respectively, between $\boldsymbol{X}^{(j)}$ and $\boldsymbol{X}$, $\boldsymbol{z}^{(j)}$ and $\boldsymbol{z}$, and $\boldsymbol{\varepsilon}^{(j)}$ and $\boldsymbol{\varepsilon}$. For every $(\boldsymbol{z}^*, \boldsymbol{X}^*) \in int(\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})})$ such that $(-\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta}, \boldsymbol{X}^*) \in int(\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})})$, we immediately have $(\boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)*}) \in int(\mathcal{S}_{(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)})})$ and $(-\boldsymbol{z}^{(j)*} - 2\boldsymbol{X}^{(j)*}\boldsymbol{\theta}, \boldsymbol{X}^{(j)*}) \in int(\mathcal{S}_{(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)})})$, as well as $d_j(\boldsymbol{\theta}; \boldsymbol{z}^{(j)*}, \boldsymbol{X}^{(j)*}) = 0$ if and only if $d_j(\boldsymbol{\theta}; \boldsymbol{z}^*, \boldsymbol{X}^*) = 0$. Therefore, we can also use the choice probability of any alternative in the choice set to achieve identification.

S.A.3 Individual Heterogeneity and Random Coefficient

Our identifying assumptions do not refer specifically to random coefficients. Here we provide sufficient conditions for our key identification assumptions I1 and I2 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, where $n = 1, ..., N$. The utility functions are then

$$u_{nj} = z_{nj} + \boldsymbol{x}'_{nj}\boldsymbol{\theta}_n + \epsilon_{nj} \text{ for } \; j = 1, \ldots, J \text{ and } u_{n0} = 0, \tag{S.A.19}$$

where $\boldsymbol{\theta}_n \in \mathcal{R}^q$ is the preference parameter vector for individual n. Now decompose the parameter vector $\boldsymbol{\theta}_n$ as $\boldsymbol{\theta}_n = \boldsymbol{\theta}^o + \boldsymbol{\delta}_n$, where $\boldsymbol{\theta}^o$ is the vector of the median of each random coefficient and

$\boldsymbol{\delta}_n = \boldsymbol{\theta}_n - \boldsymbol{\theta}^o$. Our symmetry assumption implies that $\boldsymbol{\theta}^o$ will also be the mean coefficients, as long as these means exist, but we don't impose this existence.[2]

We can rewrite the utility function (S.A.19) as $u_{nj} = z_{nj} + \boldsymbol{x}_{nj}'\boldsymbol{\theta}^o + \varepsilon_{nj}$ where $\varepsilon_{nj} = (\epsilon_{nj} + \boldsymbol{x}_{nj}'\boldsymbol{\delta}_n)$ for $j = 1, \ldots, J$. Vector $\boldsymbol{\varepsilon}_n = (\varepsilon_{n1}, \ldots, \varepsilon_{nJ})$ is often called the composite error in the presence of random coefficients. By Theorem 2.1, if this composite error vector $\boldsymbol{\varepsilon}_n$ satisfies Assumptions I1 (Exclusion Restriction) and I2 (Central Symmetry), then $\boldsymbol{\theta}^o$ is point identified under the regularity conditions given by Assumptions I3-I5. We now give sufficient conditions for I1 and I2 to hold with random coefficients.

Assumption RC.

- **RC1:** Conditional on almost every $\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{X}}$, the covariate vector $\boldsymbol{z}$ is independent of $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$, and the conditional distribution function of $\boldsymbol{z}$, $F_{\boldsymbol{z}}(\cdot \mid \boldsymbol{X})$, is absolutely continuous over its support $\mathcal{S}_{\boldsymbol{z}}(\boldsymbol{X})$.

- **RC2:** For almost every $\boldsymbol{X} \in \mathcal{S}_{\boldsymbol{X}}$, the conditional distribution function of $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$, $F_{(\boldsymbol{\epsilon}, \boldsymbol{\delta})}(\boldsymbol{t}_e, \boldsymbol{t}_c \mid \boldsymbol{X})$, where $\boldsymbol{t}_e \in \mathcal{R}^J$ and $\boldsymbol{t}_c \in \mathcal{R}^q$, admits an absolutely continuous density function, $f_{(\boldsymbol{\epsilon}, \boldsymbol{\delta})}(\boldsymbol{t}_e, \boldsymbol{t}_c \mid \boldsymbol{X})$, which is centrally symmetric about the origin, i.e.,

$$f_{(\boldsymbol{\epsilon}, \boldsymbol{\delta})}(\boldsymbol{t}_e, \boldsymbol{t}_c \mid \boldsymbol{X}) = f_{(\boldsymbol{\epsilon}, \boldsymbol{\delta})}(-\boldsymbol{t}_e, -\boldsymbol{t}_c \mid \boldsymbol{X}),$$

  for any vector $(\boldsymbol{t}_e, \boldsymbol{t}_c) \in \mathcal{S}_{(\boldsymbol{\epsilon}, \boldsymbol{\delta})}(\boldsymbol{X})$.

By the definition of the composite error vector, $\boldsymbol{\varepsilon} = \boldsymbol{\epsilon} + \boldsymbol{X}\boldsymbol{\delta}$, Assumption RC1 follows immediately from Assumption I1, because conditional independence between $\boldsymbol{z}$ and $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$ implies conditional independence between $\boldsymbol{z}$ and $\boldsymbol{\varepsilon}$.

[2] 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 $\boldsymbol{\theta}_n$ and ε_{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.

Assumption RC2 nests as a special case the random coefficients MNP model in which $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$ are assumed to be jointly normal and independent of all covariates. To show that Assumption RC2 is a sufficient condition for Assumption I2, we need to verify that the composite error vector $\boldsymbol{\varepsilon}$ satisfies conditional central symmetry, i.e.,

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

for any vector $\boldsymbol{t} \in \mathcal{S}_{\boldsymbol{\varepsilon}}(\boldsymbol{X})$. To show this, we have

$$\begin{aligned} P(\boldsymbol{\varepsilon} < \boldsymbol{t} \mid \boldsymbol{X}) &\equiv P(\boldsymbol{\epsilon} + \boldsymbol{X}\boldsymbol{\delta} < \boldsymbol{t} \mid \boldsymbol{X}) \\ &= \int_{\{(\boldsymbol{t}_e, \boldsymbol{t}_c) \mid \boldsymbol{t}_e + \boldsymbol{X}\boldsymbol{t}_c < \boldsymbol{t}\}} f_{(\boldsymbol{\epsilon}, \boldsymbol{\delta})}(\boldsymbol{t}_e, \boldsymbol{t}_c \mid \boldsymbol{X})\, d(\boldsymbol{t}_e, \boldsymbol{t}_c) \\ &= \int_{\{(\boldsymbol{t}_e^{cs}, \boldsymbol{t}_c^{cs}) \mid -\boldsymbol{t}_e^{cs} - \boldsymbol{X}\boldsymbol{t}_c^{cs} < \boldsymbol{t}\}} f_{(\boldsymbol{\epsilon}, \boldsymbol{\delta})}(-\boldsymbol{t}_e^{cs}, -\boldsymbol{t}_c^{cs} \mid \boldsymbol{X})\, d(\boldsymbol{t}_e^{cs}, \boldsymbol{t}_c^{cs}) \\ &= \int_{\{(\boldsymbol{t}_e^{cs}, \boldsymbol{t}_c^{cs}) \mid \boldsymbol{t}_e^{cs} + \boldsymbol{X}\boldsymbol{t}_c^{cs} > -\boldsymbol{t}\}} f_{(\boldsymbol{\epsilon}, \boldsymbol{\delta})}(\boldsymbol{t}_e^{cs}, \boldsymbol{t}_c^{cs} \mid \boldsymbol{X})\, d(\boldsymbol{t}_e^{cs}, \boldsymbol{t}_c^{cs}) \\ &= P(\boldsymbol{\epsilon} + \boldsymbol{X}\boldsymbol{\delta} > -\boldsymbol{t} \mid \boldsymbol{X}) \equiv P(\boldsymbol{\varepsilon} > -\boldsymbol{t} \mid \boldsymbol{X}), \end{aligned} \tag{S.A.20}$$

where the third equality in (S.A.20) holds by a change of variables (where $\boldsymbol{t}_e^{cs} = -\boldsymbol{t}_e$ and $\boldsymbol{t}_c^{cs} = -\boldsymbol{t}_c$) and the fourth equality hold by Assumption RC2. Thus we have verified that Assumption RC2 is a sufficient condition for Assumption I2.

There are four advantages of our random coefficient model and associated assumptions. First, our Assumption RC allow the joint distribution of $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$ to vary with covariates $\boldsymbol{X}$, whereas typical random coefficients models (e.g., random coefficients MNP by Hausman and Wise, 1978) assume stronger independence conditions $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp (\boldsymbol{z}, \boldsymbol{X})$, ruling out individual heterogeneity in the distribution of $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$.[3] Second, our method does not require independence between $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$, estimating these covariances, or numerical integration. In contrast, typical empirical applications of random coefficient multinomial choice estimators usually assume independence between $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ to reduce the number of parameters one must identify and estimate. 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 $\boldsymbol{\theta}_n$ that are random. Third, the econometrician is not required

[3] Even in flexible semiparametric random coefficients models like Fox and Gandhi (2016), the usual assumption is $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp (\boldsymbol{z}, \boldsymbol{X})$, ruling out the possibility that the distribution of random preferences may vary across sub-populations.

to know exactly which covariates have random coefficients and which do not. Last, our model does not require thin tails or unimodality, unlike, e.g., normal random coefficient MNP models.[4]

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.[5]) Nevertheless, we could also assume that, before normalizing, the variable z has a random coefficient, provided that the random coefficient is the same 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.

S.B A Minimum Distance Estimator and Asymptotic Properties

S.B.1 Objective Functions for Estimation

Based on the identification strategy described in Section A, we develop a minimum distance estimator (hereafter, MD estimator) for $\boldsymbol{\theta}^o \in \boldsymbol{\Theta}$ using the identifying restriction functions

$$d_j(\boldsymbol{\theta}^o; \boldsymbol{z}^{(j)*}, \mathbf{X}^{(j)*}) = 0 \tag{S.B.1}$$

[4] Each element of $\boldsymbol{\theta}^o$ is the median of the corresponding random coefficient $\boldsymbol{\theta}_n$. By Assumptions RC1 and RC2, 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.

[5] 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.

for $j = 0, \ldots, J$, where d_j is defined as the same as equation (S.A.17).

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

$$\tau_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}; \overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}\right) \equiv \varsigma_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right) \varsigma_j\left(-\boldsymbol{z}^{(j)} - 2\boldsymbol{X}^{(j)}\overline{\boldsymbol{\theta}}, \boldsymbol{X}^{(j)}\right) \varsigma_j\left(-\boldsymbol{z}^{(j)} - 2\boldsymbol{X}^{(j)}\underline{\boldsymbol{\theta}}, \boldsymbol{X}^{(j)}\right),$$

where $\boldsymbol{X}^{(j)}\overline{\boldsymbol{\theta}}$ ($\boldsymbol{X}^{(j)}\underline{\boldsymbol{\theta}}$) gives the upper (lower) bound value that the index $\boldsymbol{X}^{(j)}\boldsymbol{\theta}$ can take. A simple choice for the function $\varsigma_j(\cdot)$ is $\varsigma_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right) \equiv 1\left(|\boldsymbol{z}^{(j)}| \leq \boldsymbol{c}_1^{(j)}\right) \times 1\left(|\boldsymbol{X}^{(j)}| \leq \boldsymbol{C}_2^{(j)}\right)$, 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, $\boldsymbol{c}_1^{(j)} \in \mathcal{R}^J$ is a vector of trimming constants for the covariate vector $\boldsymbol{z}^{(j)}$, and $\boldsymbol{C}_2^{(j)} \in \mathcal{R}^{J \times q}$ is a matrix of trimming constants for the covariate matrix $\boldsymbol{X}^{(j)}$ such that $\left(\boldsymbol{c}_1^{(j)}, \boldsymbol{C}_2^{(j)}\right)$ is in the interior of the support of covariates $\mathcal{S}_{(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)})}$. Denote $\mathcal{S}_{\boldsymbol{z}^{(j)}}^{Tr}\left(\boldsymbol{X}^{(j)}, \overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}\right)$ as the largest set of values $\boldsymbol{z}^{(j)}$ given $\overline{\boldsymbol{\theta}}$, $\underline{\boldsymbol{\theta}}$, and $\boldsymbol{X}^{(j)}$, such that $\mathcal{S}_{\boldsymbol{z}^{(j)}}^{Tr}\left(\boldsymbol{X}^{(j)}, \overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}\right) \subset int\left(\mathcal{S}_{\boldsymbol{z}^{(j)}}\left(\boldsymbol{X}^{(j)}\right)\right)$. We describe the regularity conditions on the trimming function in Assumption TR.

Assumption TR. The trimming function $\tau_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}; \overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}\right)$ is strictly positive and bounded on $\mathcal{S}_{\boldsymbol{z}^{(j)}}^{Tr}\left(\boldsymbol{X}^{(j)}, \overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}\right) \times int\left(\mathcal{S}_{\boldsymbol{X}^{(j)}}\right)$, and is equal to zero on its complementary set for $j = 0, \ldots, J$.

Theorem S.B.1 *If Assumptions I and TR hold, then (i)* $Q_j(\boldsymbol{\theta}) \geq 0$ *for any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and (ii)* $Q_j(\boldsymbol{\theta}) = 0$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}^o$.

Theorem S.B.1 shows identification based on the population objective function. Proofs is available at authors' webpage.

S.B.2 MD Estimator and Regularity Conditions

Next, we derive the sample objective function based on population objective function and the asymptotic properties of the MD estimator. To ease notation, we denote the conditional means

$$E\left(y_j \mid \boldsymbol{z}^{(j)} = \boldsymbol{z}_n^{(j)}, \boldsymbol{X}^{(j)} = \boldsymbol{X}_n^{(j)}\right) \equiv \varphi_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \equiv \varphi_{j,o}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right),$$

$$E\left(y_j \mid \boldsymbol{z}^{(j)} = -\boldsymbol{z}_n^{(j)} - 2\boldsymbol{X}_n^{(j)}\boldsymbol{\theta}, \boldsymbol{X}^{(j)} = \boldsymbol{X}_n^{(j)}\right) \equiv \varphi_j\left(-\boldsymbol{z}_n^{(j)} - 2\boldsymbol{X}_n^{(j)}\boldsymbol{\theta}, \boldsymbol{X}_n^{(j)}\right) \equiv \varphi_{j,cs}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}\right),$$

and function

$$d_j(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}) \equiv \frac{\partial^J E\left(y_j \mid \boldsymbol{z}^{(j)} = \boldsymbol{z}_n^{(j)}, \boldsymbol{X}^{(j)} = \boldsymbol{X}_n^{(j)}\right)}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} - \frac{\partial^J E\left(y_j \mid \boldsymbol{z}^{(j)} = -\boldsymbol{z}_n^{(j)} - 2\boldsymbol{X}_n^{(j)}\boldsymbol{\theta}, \boldsymbol{X}^{(j)} = \boldsymbol{X}_n^{(j)}\right)}{\partial z_1^{(j)} \cdots \partial z_J^{(j)}} \tag{S.B.2}$$

$$\equiv \varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}\right).$$

where $\varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \equiv \partial^J \varphi_{j,o}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) / \partial z_1^{(j)} \cdots \partial z_J^{(j)}$ and $\varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}\right)$ is defined in the similar way as $\varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)$. Now, consider a leave-one-out (LOO) Nadaraya-Watson (NW) estimator for $\varphi_{j,o,-n}^{(J)}$ as $\hat{\varphi}_{j,o,-n}^{(J)}(\cdot,\cdot) = \frac{\frac{1}{N-1}\sum_{m=1,m\neq n}^{N} y_{mj} K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)} - \cdot\right) K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)} - \cdot\right)}{\frac{1}{N-1}\sum_{m=1,m\neq n}^{N} K_{\boldsymbol{h_z}}\left(\boldsymbol{z}_m^{(j)} - \cdot\right) K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)} - \cdot\right)}$, where $K_{\boldsymbol{h_z}}\left(\boldsymbol{z}_m^{(j)} - \cdot\right) = \prod_{l=1}^{J} h_{z_l}^{-1} k\left(h_{z_l}^{-1}\left(z_{ml}^{(j)} - \cdot\right)\right)$, and $K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)} - \cdot\right) = \prod_{l=1}^{J}\prod_{r=1}^{q} h_{x_{lr}}^{-1} k\left(h_{x_{lr}}^{-1}\left(x_{mlr}^{(j)} - \cdot\right)\right)$. The properties of the kernel function k and those of the bandwidth $\boldsymbol{h_z} \equiv (h_{z_1}, \cdots, h_{z_J})'$ and $\boldsymbol{h_X} \equiv (h_{x_{1,1}}, \cdots, h_{x_{1,q}}, \cdots, h_{x_{J,1}}, \cdots, h_{x_{J,q}})'$ are defined in Assumptions E3 and E4 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.

Now we can construct the estimator for function $d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)$ in (S.B.2) as

$$\hat{d}_{j,-n}\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \equiv \hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}\right). \tag{S.B.3}$$

By replacing the expectation in $Q_j(\boldsymbol{\theta})$ with its sample mean and function $d_j(\boldsymbol{\theta};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)})$ with its LOO estimator $\hat{d}_{j,-n}\left(\boldsymbol{\theta};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)$, we define the MD estimator

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} Q_{Nj}(\boldsymbol{\theta}),$$

$$\text{where} \qquad Q_{Nj}(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{n=1}^{N}\left[\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\hat{d}_{j,-n}\left(\boldsymbol{\theta};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right]^2.$$

We denote the gradient of the objective function as $\boldsymbol{q}_{Nj}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}Q_{Nj}(\boldsymbol{\theta})$ and the Hessian matrix of the objective function as $\boldsymbol{H}_{Nj}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'}Q_{Nj}(\boldsymbol{\theta})$. The smoothness of the objective function suggests the first-order condition (FOC): $\boldsymbol{q}_{Nj}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{0}_q$. Applying the standard first-order Taylor expansion to $\boldsymbol{q}_{Nj}\left(\hat{\boldsymbol{\theta}}\right)$ around the true parameter vector $\boldsymbol{\theta}^o$ yields $\boldsymbol{q}_{Nj}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{q}_{Nj}(\boldsymbol{\theta}^o) + \boldsymbol{H}_{Nj}\left(\tilde{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^o\right)$, where $\tilde{\boldsymbol{\theta}}$ is a vector between the MD estimator $\hat{\boldsymbol{\theta}}$ and the true parameter vector $\boldsymbol{\theta}^o$. Then the influence function will be given by

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o = -\left[\boldsymbol{H}_{Nj}\left(\tilde{\boldsymbol{\theta}}\right)\right]^{-1}\boldsymbol{q}_{Nj}(\boldsymbol{\theta}^o). \tag{S.B.4}$$

We will show that $\boldsymbol{H}_{Nj}\left(\tilde{\boldsymbol{\theta}}\right) \to_p \boldsymbol{H}_j(\boldsymbol{\theta}^o)$, where

$$\boldsymbol{H}_j(\boldsymbol{\theta}^o) = E\left\{\tau_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\nabla_{\boldsymbol{\theta}}d_j\left(\boldsymbol{\theta}^o;\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left[\nabla_{\boldsymbol{\theta}}d_j\left(\boldsymbol{\theta}^o;\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right]'\right\}, \tag{S.B.5}$$

and $\sqrt{N}\boldsymbol{q}_{Nj}(\boldsymbol{\theta}^o) \to_d N(\mathbf{0}_q,\boldsymbol{\Omega}_j)$, where $\boldsymbol{\Omega}_j$ is the probability limit of the variance-covariance matrix of $\boldsymbol{q}_{Nj}(\boldsymbol{\theta}^o)$. To obtain these properties, we assume the following regularity conditions.

Assumption E.

- **E1:** $\{(\boldsymbol{y}_n,\boldsymbol{z}_n,\boldsymbol{X}_n)\,,\text{ for } n=1,\ldots,N\}$ is a random sample drawn from the infinite population distribution.

- **E2:** The following smoothness conditions hold: (a) The density function $f_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)$ is continuous in the components of $\boldsymbol{z}^{(j)}$ for all $\boldsymbol{z}^{(j)} \in \mathcal{S}_{\boldsymbol{z}}^{Tr}\left(\boldsymbol{X}^{(j)}, \overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}\right)$ and $\boldsymbol{X}^{(j)} \in int\left(\mathcal{S}_{\boldsymbol{X}}\right)$. In addition, $f_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)$ is bounded away from zero uniformly over its support. (b) Functions $f_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)$, $g_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)$ and $\varphi_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)$ are s ($s \geq J+1$) times continuously differentiable in the components of $\boldsymbol{z}^{(j)}$ for all $\boldsymbol{z}^{(j)} \in \mathcal{S}_{\boldsymbol{z}}^{Tr}\left(\boldsymbol{X}^{(j)}, \overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}\right)$ and have bounded derivatives.

- **E3:** 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) $\sup_u \left|k^{(1)}(u)\right|^2 < \infty$, where $k^{(1)}(u)$ is the first derivative of $k(u)$.

- **E4:** The bandwidth vector $\boldsymbol{h}_{\boldsymbol{z}} \equiv (h_{z_1}, \cdots, h_{z_J})' = (h_N, \cdots, h_N)'$ is a $J \times 1$ vector and the bandwidth $\boldsymbol{h}_{\boldsymbol{X}} \equiv (h_{x_{1,1}}, \cdots, h_{x_{1,q}}, \cdots, h_{x_{J,1}}, \cdots, h_{x_{J,q}})' = (h_N, \cdots, h_N, \cdots, h_N, \cdots, h_N)'$ 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) $\sqrt{N} h_N^{2s} \to 0$ and $\sqrt{N}(\ln N)\left(Nh_N^{2(J+1)+J+Jq}\right)^{-1} \to 0$ as $N \to \infty$.

- **E5:** The components of the random vectors $\partial^J \varphi_{j,o} / \partial z_1^{(j)} \cdots \partial z_J^{(j)}, \partial^J \varphi_{j,cs} / \partial z_1^{(j)} \cdots \partial z_J^{(j)}$ and the random matrix $\left(\partial^J \boldsymbol{\xi}_j / \partial z_1^{(j)} \cdots \partial z_J^{(j)}\right)\left[y^{(j)}, \boldsymbol{z}^{(j)\prime}\right]$ have finite second moments. Also, $\partial^J \varphi_{j,o} / \partial z_1^{(j)} \cdots \partial z_J^{(j)}, \partial^J \varphi_{j,cs} / \partial z_1^{(j)} \cdots \partial z_J^{(j)}$ and $\partial^J \boldsymbol{\xi}_j \varphi_{j,o} / \partial z_1^{(j)} \cdots \partial z_J^{(j)}$, $\partial^J \boldsymbol{\xi}_j \varphi_{j,cs} / \partial z_1^{(j)} \cdots \partial z_J^{(j)}$ where $\boldsymbol{\xi}_j\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}, \boldsymbol{\theta}^o\right) = \tau_j^2\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right) \nabla_{\boldsymbol{\theta}} d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)$, satisfy the following Lipschitz conditions: for some $m\left(\boldsymbol{z}^{(j)},\cdot\right)$

$$\left|\frac{\partial^J\varphi_{j,o}\left(\boldsymbol{z}^{(j)}+\boldsymbol{t},\cdot\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}-\frac{\partial^J\varphi_{j,o}\left(\boldsymbol{z}^{(j)},\cdot\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right|<m\left(\boldsymbol{z}^{(j)},\cdot\right)\|\boldsymbol{t}\|;\left|\frac{\partial^J\varphi_{j,cs}\left(\boldsymbol{z}^{(j)}+\boldsymbol{t},\cdot\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}-\frac{\partial^J\varphi_{j,cs}\left(\boldsymbol{z}^{(j)},\cdot\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right|<m\left(\boldsymbol{z}^{(j)},\cdot\right)\|\boldsymbol{t}\|;$$
$$\left|\frac{\partial^J\boldsymbol{\xi}_j\varphi_{j,o}\left(\boldsymbol{z}^{(j)}+\boldsymbol{t},\cdot\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}-\frac{\partial^J\boldsymbol{\xi}_j\varphi_{j,o}\left(\boldsymbol{z}^{(j)},\cdot\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right|<m\left(\boldsymbol{z}^{(j)},\cdot\right)\|\boldsymbol{t}\|;\left|\frac{\partial^J\boldsymbol{\xi}_j\varphi_{j,cs}\left(\boldsymbol{z}^{(j)}+\boldsymbol{t},\cdot\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}-\frac{\partial^J\boldsymbol{\xi}_j\varphi_{j,cs}\left(\boldsymbol{z}^{(j)},\cdot\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right|<m\left(\boldsymbol{z}^{(j)},\cdot\right)\|\boldsymbol{t}\|;$$

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

- **E6:** The matrix $\boldsymbol{H}_j$ defined by (S.B.5) is nonsingular and positive definite.

Assumption E2 gives the smoothness condition of the density and the choice probability. Assumption E3 collects restrictions for the kernel function. Assumption E4 describes the conditions on the bandwidth to achieve $\sqrt{N}$ asymptotics. Assumption E5 imposes standard bounded moment and dominance conditions. Assumption E6 requires the hessian matrix is strictly positive definite. Given these regularity assumptions, we will show asymptotic properties of the estimator.

S.B.3 Asymptotic Properties

The next two theorems establish the asymptotic properties of the MD estimator. The proofs are available on authors' webpage.

Theorem S.B.2 *If Assumptions I, TR, E1-E5 hold, then the MD estimator $\hat{\boldsymbol{\theta}}$ converges to the true parameter vector $\boldsymbol{\theta}^o\in\boldsymbol{\Theta}$ in probability.*

Theorem S.B.3 *Let Assumptions I, TR and E hold. Then*

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

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^o\right)=-N^{-1/2}\sum_{n=1}^{N}\boldsymbol{H}_j^{-1}\boldsymbol{t}_{nj}+o_p(1),$$

where $\boldsymbol{t}_{nj}\equiv(\upsilon_{nj,o}-\upsilon_{nj,cs})\,\partial^J\left[\tau_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\nabla_{\boldsymbol{\theta}}d_j\left(\boldsymbol{\theta}^o;\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right]/\partial z_1^{(j)}\ldots\partial z_J^{(j)}$ *with scalars* $\upsilon_{nj,o}\equiv y_{nj}-\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)$ *and* $\upsilon_{nj,cs}\equiv y_{nj}-\varphi_{j,cs}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)$.

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

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{o}\right)\rightarrow_{d} N\left(\mathbf{0}_{q}, \boldsymbol{H}_{j}^{-1}\boldsymbol{\Omega}_{j}\boldsymbol{H}_{j}^{-1}\right)$$

where matrix $\boldsymbol{\Omega}_j \equiv E\left(\boldsymbol{t}_{nj}\boldsymbol{t}_{nj}'\right)$ *and* $\boldsymbol{H}_j$ *is defined by (S.B.5).*

In proofs, we show that the leading term of the numerator of this estimator in equation (S.B.5) 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.

Our simplest MD estimator only requires observing a single choice j (e.g., selecting the outside option or not) and minimizing $Q_{Nj}(\boldsymbol{\theta})$, 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 $Q_{Nj}(\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$, 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 $Q_{Nj}(\boldsymbol{\theta})$ 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.

S.C Proof of Identification

Proof of Theorem S.A.1: First, we show that $\mathcal{D}_0(\boldsymbol{\theta}^o)$ is a set of measure zero. If not, assume that there is a point $(\boldsymbol{z}^*, \boldsymbol{X}^*)$ in set $\mathcal{D}_0(\boldsymbol{\theta}^o)$. By definition in equation (8), both points $(\boldsymbol{z}^*, \boldsymbol{X}^*)$ and $(-\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta}^o, \boldsymbol{X}^*)$ are in set $int(\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})})$. By Assumptions I1, I2, and equations (5)-(7), we have function

$$d_0(\boldsymbol{\theta}^o; \boldsymbol{z}^*, \boldsymbol{X}^*) = (-1)^J \left[f_{\boldsymbol{\varepsilon}}\left(-\boldsymbol{z}^* - \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{X} = \boldsymbol{X}^*\right) - f_{\boldsymbol{\varepsilon}}\left(\boldsymbol{z}^* + \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{X} = \boldsymbol{X}^*\right)\right] = 0,$$

which is a contradiction with definition in equation (8).

Next, we prove that $\Pr[\,(\boldsymbol{z}^*, \boldsymbol{X}^*) \in \mathcal{D}_0(\boldsymbol{\theta})\,] > 0$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}^o$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and parameter space $\boldsymbol{\Theta}$ satisfies Assumption I3. Denote the set $\mathcal{X}(\boldsymbol{\theta}) \equiv \{\boldsymbol{X}^* \in \mathcal{S}_{\boldsymbol{X}} \mid \boldsymbol{X}^*(\boldsymbol{\theta} - \boldsymbol{\theta}^o) \neq \mathbf{0}\}$, which is a collection of covariate values at which $\boldsymbol{X}\boldsymbol{\theta} \neq \boldsymbol{X}\boldsymbol{\theta}^o$. By Assumption I4(a) and the fact $\boldsymbol{\theta} - \boldsymbol{\theta}^o \neq \mathbf{0}_q$, $\mathcal{X}(\boldsymbol{\theta})$ is a subset in the support of $\mathcal{S}_{\boldsymbol{X}}$ with positive measure, that is,

$$\Pr\left[\boldsymbol{X}^* \in \mathcal{X}(\boldsymbol{\theta})\right] > 0. \tag{S.C.1}$$

Recall that we use $\boldsymbol{X}_c$ and $\boldsymbol{X}_d$, respectively, to denote the continuous and discrete covariates in $\boldsymbol{X}$. We define the interior of the support of $\boldsymbol{X}$ as $int\,(\mathcal{S}_{\boldsymbol{X}}) \equiv$

$\left\{(\boldsymbol{X}_c^*, \boldsymbol{X}_d^*) \in \mathcal{S}_{(\boldsymbol{X}_c, \boldsymbol{X}_d)} \mid \boldsymbol{X}_c^* \in int\,(\mathcal{S}_{\boldsymbol{X}_c}(\boldsymbol{X}_d^*))\,, \boldsymbol{X}_d^* \in \mathcal{S}_{\boldsymbol{X}_d}\right\}$. Define

$$\widetilde{\mathcal{S}}_{(\boldsymbol{z},\boldsymbol{X})}(\boldsymbol{\theta}) \;\; \equiv \;\; \left\{ (\boldsymbol{z}^*,\; \boldsymbol{X}^*) \in \mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})} \;\middle|\; \boldsymbol{z}^* \in \widetilde{\mathcal{S}}_{\boldsymbol{z}}(\boldsymbol{X}^*),\; \boldsymbol{X}^* \in \mathcal{X}(\boldsymbol{\theta}) \cap int\,(\mathcal{S}_{\boldsymbol{X}}) \right\}, \tag{S.C.2}$$

where $\widetilde{\mathcal{S}}_{\boldsymbol{z}}(\boldsymbol{X}^*)$ satisfies Assumption I4(c). By construction, set $\widetilde{\mathcal{S}}_{(\boldsymbol{z},\boldsymbol{X})}(\boldsymbol{\theta})$ is a Lebesgue measurable subset of $int(\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})})$. Next we construct a subset in the support of covariates $(\boldsymbol{z},\boldsymbol{X})$ as follows:

$$\widetilde{\mathcal{D}}_0(\boldsymbol{\theta}) \equiv \left\{(\boldsymbol{z}^*,\boldsymbol{X}^*) \in \widetilde{\mathcal{S}}_{(\boldsymbol{z},\boldsymbol{X})}(\boldsymbol{\theta}) \mid d_0(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*) \neq 0\right\} \tag{S.C.3}$$

which is also a subset of $\mathcal{D}_0(\boldsymbol{\theta})$ because $(-\boldsymbol{z}^* - 2\boldsymbol{X}^*\boldsymbol{\theta},\boldsymbol{X}^*) \in int(\mathcal{S}_{(\boldsymbol{z},\boldsymbol{X})})$ for any $(\boldsymbol{z}^*,\boldsymbol{X}^*) \in \widetilde{\mathcal{S}}_{(\boldsymbol{z},\boldsymbol{X})}(\boldsymbol{\theta})$. Under Assumptions I1-I4 and I5(a), both sets $\mathcal{D}_0(\boldsymbol{\theta})$ and $\widetilde{\mathcal{D}}_0(\boldsymbol{\theta})$ are Lebesgue measurable. Theorem S.A.1 is proved if we show $P[(\boldsymbol{z}^*,\boldsymbol{X}^*) \in \widetilde{\mathcal{D}}_0(\boldsymbol{\theta})] > 0$ since $\widetilde{\mathcal{D}}_0(\boldsymbol{\theta}) \subseteq \mathcal{D}_0(\boldsymbol{\theta})$. Now

$$\begin{aligned}\Pr\left[(\boldsymbol{z}^*,\ \boldsymbol{X}^*) \in \widetilde{\mathcal{S}}_{(\boldsymbol{z},\boldsymbol{X})}(\boldsymbol{\theta})\right] &= \Pr\left[\boldsymbol{X}^* \in \mathcal{X}(\boldsymbol{\theta}) \cap int\left(\mathcal{S}_{\boldsymbol{X}}\right)\right] \\ &\quad \times \Pr\left[\boldsymbol{z}^* \in \widetilde{\mathcal{S}}_{\boldsymbol{z}}(\boldsymbol{X}^*) \mid \boldsymbol{X}^* \in \mathcal{X}(\boldsymbol{\theta}) \cap int\left(\mathcal{S}_{\boldsymbol{X}}\right)\right],\end{aligned} \tag{S.C.4}$$

where the first probability on the right of equation (S.C.4) is positive by (S.C.1), and the second is positive by Assumption I4(c). Under Assumptions I1, I2, and equations (5)-(7), we have

$$d_0(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*) = (-1)^J\left[f_{\boldsymbol{\varepsilon}}\left(-\boldsymbol{z}^* - \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{X} = \boldsymbol{X}^*\right) - f_{\boldsymbol{\varepsilon}}\left(\boldsymbol{z}^* + 2\boldsymbol{X}^*\boldsymbol{\theta} - \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{X} = \boldsymbol{X}^*\right)\right]$$

for every $(\boldsymbol{z}^*,\boldsymbol{X}^*) \in \widetilde{\mathcal{S}}_{(\boldsymbol{z},\boldsymbol{X})}(\boldsymbol{\theta})$. Define $\boldsymbol{r} = 2\boldsymbol{X}^*(\boldsymbol{\theta} - \boldsymbol{\theta}^o)$. Then $\boldsymbol{r} \neq \boldsymbol{0}_J$ because $\boldsymbol{X}^* \in \mathcal{X}(\boldsymbol{\theta})$. We can write $\boldsymbol{z}^* + 2\boldsymbol{X}^*\boldsymbol{\theta} - \boldsymbol{X}^*\boldsymbol{\theta}^o = \boldsymbol{r} + \boldsymbol{z}^* + \boldsymbol{X}^*\boldsymbol{\theta}^o$ so function

$$d_0(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*) = (-1)^J\left[f_{\boldsymbol{\varepsilon}}\left(-\boldsymbol{z}^* - \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{X} = \boldsymbol{X}^*\right) - f_{\boldsymbol{\varepsilon}}\left(\boldsymbol{r} + \boldsymbol{z}^* + \boldsymbol{X}^*\boldsymbol{\theta}^o \mid \boldsymbol{X} = \boldsymbol{X}^*\right)\right]. \tag{S.C.5}$$

We claim that

$$\Pr\left(d_0(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*) \neq 0 \;\middle|\; (\boldsymbol{z}^*,\boldsymbol{X}^*) \in \widetilde{\mathcal{S}}_{(\boldsymbol{z},\boldsymbol{X})}(\boldsymbol{\theta})\right) > 0. \tag{S.C.6}$$

If (S.C.6) is not true, then for almost every $(\boldsymbol{z}^*,\boldsymbol{X}^*) \in \widetilde{\mathcal{S}}_{(\boldsymbol{z},\boldsymbol{X})}(\boldsymbol{\theta})$ we get $d_0(\boldsymbol{\theta};\boldsymbol{z}^*,\boldsymbol{X}^*) = 0$, which implies that $f_{\boldsymbol{\varepsilon}}(\boldsymbol{t} \mid \boldsymbol{X} = \boldsymbol{X}^*) = f_{\boldsymbol{\varepsilon}}(\boldsymbol{r} - \boldsymbol{t} \mid \boldsymbol{X} = \boldsymbol{X}^*)$ for every $\boldsymbol{t} \in \widetilde{\mathcal{S}}_{\boldsymbol{\varepsilon}}(\boldsymbol{X}^*)$ and $\boldsymbol{r} - \boldsymbol{t} \in \widetilde{\mathcal{S}}_{\boldsymbol{\varepsilon}}(\boldsymbol{X}^*)$ by (S.C.5) and Assumption I5(a). This is possible only if $\boldsymbol{r} = \boldsymbol{0}_J$ by Assumption I5(b), which contradicts $\boldsymbol{r} \neq \boldsymbol{0}_J$. We have therefore proved that $\Pr[(\boldsymbol{z}^*,\boldsymbol{X}^*) \in \widetilde{\mathcal{D}}_0(\boldsymbol{\theta})] > 0$ by (S.C.3), (S.C.4), and (S.C.6). *Q.E.D.*

S.D Monte Carlo Details

As discussed in the paper, our Monte Carlo design includes 4 data generating processes (DGPs). Details of the distribution of each DGP are provided in Table 1.

Table 1: Designs of the Data Generating Processes (DGPs)

DGP	Distribution of θ_n	Distribution of ε_{nj}
1	$\theta_n = 0.2$	$\varepsilon_{nj} = \epsilon_{nj}$
2	$\theta_n = 0.2$	$\varepsilon_{nj} = \frac{1}{2} e^{2x_{nj}} \epsilon_{nj}$,
3	$\theta_n = 0.2 + \delta_n$ where $\delta_n = \frac{1}{2}\vartheta_n$	$\varepsilon_{nj} = \frac{1}{2}\epsilon_{nj}$
4	$\theta_n = 0.2 + \delta_n$ where $\delta_n = (e^{x_{n1}} + e^{x_{n2}}) \times \vartheta_n$	$\varepsilon_{nj} = \frac{1}{2}\epsilon_{nj}$

Note: both ϑ_n and ϵ_{nj} are standard normal random varaibles, and they are independent of each other and all the covariates, and i.i.d. across the subscripted dimension(s).

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 $Q_{N0}(\boldsymbol{\theta})$, and the case where the researcher also observes which alternative is chosen by each decision maker, and so minimizes the sum of $Q_{Nj}(\boldsymbol{\theta})$ 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 are independent, identically distributed 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 $h_j = sd(z_{nj})N^{(-1/22)}$, where $j = 1, 2$. Our bandwidth is derived by minimizing the mean squared errors (MSE) of the second order derivative of the choice probability. 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 would be an alternative way to select the bandwidth, but is more computation-intensive.

Additional References

Bliemer, M. C.J. and Rose J. M. (2013): "Confidence Intervals of Willingness-to-Pay for Random Coefficient Logit Models," *Transportation Research Part B: Methodological*, 58, 199-214.

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

Imbens, G. W. and Ridder G. (2009): "Estimation and Inference for Generalized Full and Partial Means and Average Derivatives," Harvard Univesity, Working Paper.

Lewbel, A. (2019): "The Identification Zoo: Meanings of Identification in Econometrics," *Journal of Economic Literature*, 57, 835-903.

Newey, W.K. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10(2), 233-253.

Newey, W.K. and McFadden, D. (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, Vol. 4, 2111-2245.

Powell, J.L., Stock, J., and Stocker, T. (1989): "Semiparametric Estimation of Index Models," *Econometrica*, 57, 1403-1430.

Sherman, R.P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123-137.

Sherman, R.P. (1994): "U-process in the Analysis of a Generalized Semiparametric Regression Estimator," *Econometric Theory*, 10(2), 372-395.

Online 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 December 2021

S.D Proofs Regarding Estimation

In this section, we provide the proofs of Theorems S.B.1-S.B.3 in Section S.B of the Supplementary Appendix and their related lemmas. Specifically, Section S.D.1 provides the proof of Theorem S.B.1 on the population sample objective function; Section S.D.2 collects preliminary lemmas needed for the asympotic properties of the MD estimator defined in Section S.B.2; Section S.D.3 provides the proofs of Theorem S.B.2, the consistency of the MD estimator; and Section S.D.4 gives the proofs of Theorem S.B.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.

§Economics, New York University Shanghai; Shanghai; email: amanda.yu.zhou@nyu.edu

S.D.1 Proof of the Population Objective Function

Proof of Theorem S.B.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

$$
\begin{aligned}
Q_j(\boldsymbol{\theta}) &\equiv \frac{1}{2}\mathbb{E}\left[\tau_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) d_j\left(\boldsymbol{\theta}; z_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right]^2 \\
&= \frac{1}{2}\mathbb{E}\left\{\mathbb{E}\left[\tau_j^2\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) d_j^2\left(\boldsymbol{\theta}; z_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\middle| \boldsymbol{X}_n^{(j)}\right]\right\}
\end{aligned}
$$

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

$$
\mathbb{E}\left[\tau_j^2\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) d_j^2\left(\boldsymbol{\theta}; z_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\middle| \boldsymbol{X}_n^{(j)}\right] = 0 \tag{S.D.1}
$$

when $\boldsymbol{\theta} = \boldsymbol{\theta}^o$. The equality in (S.D.1) holds because, conditional on $\boldsymbol{X}_n^{(j)}$, when $\tau_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) > 0$, $d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) = 0$; and in addition, when $\tau_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) = 0$, the product term in the expectation is also equal to zero. Combining these parts gives the desired existence.

To show the uniqueness, consider any $\boldsymbol{\theta}$ in the parameter space such that $\boldsymbol{\theta} \neq \boldsymbol{\theta}^o$. We have

$$
\begin{aligned}
&Q_j(\boldsymbol{\theta}) - Q_j(\boldsymbol{\theta}^o) \qquad \text{(S.D.2)}\\
&= \frac{1}{2}\mathbb{E}\left[\tau_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right]^2 - \frac{1}{2}\mathbb{E}\left[\tau_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right]^2 \\
&= \frac{1}{2}\mathbb{E}\left[\tau_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\left(d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right)\right]^2 \\
&\quad + \mathbb{E}\left[\tau_j^2\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\left(d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right)\right] > 0.
\end{aligned}
$$

The last inequality in (S.D.2) holds because the first term on its right-hand side is strictly positive, as there exists some $\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}$ such that $\tau_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) > 0$ and $d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \neq 0$ when $\boldsymbol{\theta} \neq \boldsymbol{\theta}^o$ by Theorem S.A.1 and the identification results in Supplementary Appendix Section S.A.1; and the second term equals to zero since $d_j\left(\boldsymbol{\theta}^o;\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)=0$.

Q.E.D.

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

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

Lemma S.D.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*

$$U_N \equiv \frac{1}{N(N-1)}\sum\nolimits_{m=1,m\neq n}^{N}\sum\nolimits_{n=1}^{N}\psi_N(\omega_m,\omega_n)$$

Define

$$\hat{U}_N \equiv \mu_N + \frac{1}{N}\sum\nolimits_{m=1}^{N}\left(r_{N1}(\omega_m)-\mu_N\right)+\frac{1}{N}\sum\nolimits_{n=1}^{N}\left(r_{N2}(\omega_n)-\mu_N\right)$$

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

Lemma S.D.2 *Under Assumptions E2-E5,*

$$\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\hat{f}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-f_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|=O_p\left(\sqrt{\frac{\ln N}{Nh_N^{J+Jq}}}+h_N^s\right)=o_p\left(N^{-1/4}\right)$$

$$\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\hat{g}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-g_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|=O_p\left(\sqrt{\frac{\ln N}{Nh_N^{J+Jq}}}+h_N^s\right)=o_p\left(N^{-1/4}\right)$$

$$\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\hat{\varphi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\varphi_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|=O_p\left(\sqrt{\frac{\ln N}{Nh_N^{J+Jq}}}+h_N^s\right)=o_p\left(N^{-1/4}\right)$$

Proof of Lemma S.D.2: The proofs for three terms are similar. We wil focus on the proof for $\hat{g}_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)$. 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 Asssumption 10 using Lemma 8.10 in Newey and McFadden (1994). *Q.E.D.*

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

Lemma S.D.3 *Under Assumptions E2-E5, for $t = 1, \ldots, J$,*

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

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

$$\sup_{\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \in \mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)}} \left|\hat{\varphi}_j^{(t)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \varphi_j^{(t)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right| = O_p\left(\sqrt{\frac{\ln N}{N h_N^{J+2t+Jq}}} + h_N^s\right) = o_p\left(N^{-1/4}\right)$$

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

Lemma S.D.4 *Under Assumptions E2-E5, for $t = 1, \ldots, J$,*

$$\sup_{\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \in \mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)}} \left\| \nabla_{\boldsymbol{\theta}} \left(\hat{f}_j^{(t)} \left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)} \right) \right) - \nabla_{\boldsymbol{\theta}} \left(f_j^{(t)} \left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)} \right) \right) \right\|$$

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

$$\sup_{\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \in \mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)}} \left\| \nabla_{\boldsymbol{\theta}} \left(\hat{g}_j^{(t)} \left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)} \right) \right) - \nabla_{\boldsymbol{\theta}} \left(g_j^{(t)} \left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)} \right) \right) \right\|$$

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

$$\sup_{\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \in \mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)}} \left\| \nabla_{\boldsymbol{\theta}} \left(\hat{\varphi}_j^{(t)} \left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)} \right) \right) - \nabla_{\boldsymbol{\theta}} \left(\varphi_j^{(t)} \left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)} \right) \right) \right\|$$

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

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

S.D.3 Consistency of the MD Estimator

Proof of Theorem S.B.1: We apply Theorem S.A.1 of Newey and McFadden (1994) to show the consistency of the MD estimator. Theorem S.A.1 in Newey and McFadden (1994) requires the following fours conditions: (1) the population objective function $Q_j(\boldsymbol{\theta})$ is uniquely minimized at $\boldsymbol{\theta}^o \in \boldsymbol{\Theta}$; (2) the parameter space $\boldsymbol{\Theta}$ is compact; (3) the population objective function $Q_j(\boldsymbol{\theta})$ is continuous; and (4) the sample objective function $Q_{Nj}(\boldsymbol{\theta})$ converges uniformly in probability to $Q_j(\boldsymbol{\theta})$ over the parameter space.

Our Theorem S.B.1 directly implies Condition (1). Condition (2) follows from Assumption E3. 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 $\bar{Q}_{Nj}(\boldsymbol{\theta})$, defined as

$$\bar{Q}_{Nj}(\boldsymbol{\theta}) = \frac{1}{2N}\sum\nolimits_{n=1}^{N}\left[\tau_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right]^2.$$

Following the triangle inequality, we have

$$\left|Q_{Nj}(\boldsymbol{\theta}) - Q_j(\boldsymbol{\theta})\right| \leq \left|Q_{Nj}(\boldsymbol{\theta}) - \bar{Q}_{Nj}(\boldsymbol{\theta})\right| + \left|\bar{Q}_{Nj}(\boldsymbol{\theta}) - Q_j(\boldsymbol{\theta})\right|. \tag{S.D.3}$$

Then, it is sufficient to show that the two terms on the right side of (S.D.3) go to zero uniformly, that is, (i) $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|Q_{Nj}(\boldsymbol{\theta}) - \bar{Q}_{Nj}(\boldsymbol{\theta})\right| = o_p(1)$ and (ii) $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\bar{Q}_{Nj}(\boldsymbol{\theta}) - Q_j(\boldsymbol{\theta})\right| = o_p(1)$.

For Part (i), we observe that

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|Q_{Nj}(\boldsymbol{\theta}) - \bar{Q}_{Nj}(\boldsymbol{\theta})\right| \qquad\qquad \text{(S.D.4)}\\
&= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{2N}\sum\nolimits_{n=1}^{N}\left[\tau_j^2\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\left(\hat{d}_{j,-n}^2\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - d_j^2\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right)\right]\right|\\
&= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{2N}\sum\nolimits_{n=1}^{N}\tau_j^2\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\left(\hat{d}_{j,-n}\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) + d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right)\right.\\
&\qquad \left.\times\left(\hat{d}_{j,-n}\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right)\right|\\
&\leq C\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\sup_{\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)}, \boldsymbol{X}^{(j)}\right)}}\left|\hat{d}_{j,-n}\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - d_j\left(\boldsymbol{\theta}; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right| = o_p(1).
\end{aligned}$$

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

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

is bounded by the product of a constant and the derivative functions shown by Lemma S.D.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 $\bar{Q}_{Nj}(\boldsymbol{\theta})$ to $Q_j(\boldsymbol{\theta})$. Next we can conclude the uniformity by showing stochastic equicontinuity, that is,

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

Following Andrews (1994), the stochastic equicontinuity can be shown by verifying that $\bar{Q}_{Nj}(\boldsymbol{\theta})$ is the type II class of function, satisfying the Lipschitz condition $\left|\bar{Q}_{Nj}\left(\boldsymbol{\theta}^{(1)}\right)-\bar{Q}_{Nj}\left(\boldsymbol{\theta}^{(2)}\right)\right|\leq C||\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}^{(2)}||$. 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.D.4 Asymptotic Linearity and Normality of the MD Estimator

In this section, we first show the lemmas that contribute to the proof of Theorem S.B.3

Lemma S.D.5 *Under Assumptions I,TR and E,* $\boldsymbol{H}_{Nj}\left(\tilde{\boldsymbol{\theta}}\right)\rightarrow_p\boldsymbol{H}_j$*, where*

$$\boldsymbol{H}_j=\mathbb{E}\left\{\tau_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\nabla_{\boldsymbol{\theta}}d_j\left(\boldsymbol{\theta}^o;\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left[\nabla_{\boldsymbol{\theta}}d_j\left(\boldsymbol{\theta}^o;\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right]'\right\}$$

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

(i) $\boldsymbol{H}_{Nj}\left(\tilde{\boldsymbol{\theta}}\right)=\boldsymbol{H}_{Nj,1}\left(\tilde{\boldsymbol{\theta}}\right)+\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)$, where

$$\boldsymbol{H}_{Nj,1}\left(\tilde{\boldsymbol{\theta}}\right)=\frac{1}{N}\sum_{n=1}^N\tau_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\hat{d}_{j,-n}\left(\tilde{\boldsymbol{\theta}};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'}\hat{d}_{j,-n}\left(\tilde{\boldsymbol{\theta}};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right],$$

$$\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)=\frac{1}{N}\sum_{n=1}^N\tau_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\nabla_{\boldsymbol{\theta}}\hat{d}_{j,-n}\left(\tilde{\boldsymbol{\theta}};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left[\nabla_{\boldsymbol{\theta}}\hat{d}_{j,-n}\left(\tilde{\boldsymbol{\theta}};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right]',$$

(ii) $\boldsymbol{H}_{Nj,1}\left(\tilde{\boldsymbol{\theta}}\right)=o_p(1)$, and (iii) $\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)\rightarrow_p\boldsymbol{H}_j$.

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

$$\boldsymbol{H}_{Nj,1}\left(\tilde{\boldsymbol{\theta}}\right)=\left[\boldsymbol{H}_{Nj,1}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{Nj,1}\left(\boldsymbol{\theta}^{o}\right)\right]+\boldsymbol{H}_{Nj,1}\left(\boldsymbol{\theta}^{o}\right)=o_{p}\left(1\right)$$

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

$$\tau_{j}^{2}\left(\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)d_{j}\left(\boldsymbol{\theta}^{o};\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)\left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'}d_{j}\left(\boldsymbol{\theta}^{o};\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)\right]'=\mathbf{0}_{q\times q}$$

and

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\sup_{\left(\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\hat{d}_{j,-n}\left(\boldsymbol{\theta};\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)-d_{j}\left(\boldsymbol{\theta};\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)\right|$$

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

For Part (iii), define

$$\boldsymbol{H}_{j}\left(\boldsymbol{\theta}\right)=\mathbb{E}\left\{\tau_{j}^{2}\left(\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)\nabla_{\boldsymbol{\theta}}d_{j}\left(\boldsymbol{\theta};\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)\left[\nabla_{\boldsymbol{\theta}}d_{j}\left(\boldsymbol{\theta};\boldsymbol{z}_{n}^{(j)},\boldsymbol{X}_{n}^{(j)}\right)\right]'\right\}$$

and we have

$$\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{j}\left(\boldsymbol{\theta}^{o}\right)=\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{j}\left(\boldsymbol{\theta}^{o}\right)=\left[\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{j}\left(\tilde{\boldsymbol{\theta}}\right)\right]+\left[\boldsymbol{H}_{j}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{j}\left(\boldsymbol{\theta}^{o}\right)\right].$$

By the triangle inequality theorem, we have that

$$\left\|\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{j}\left(\boldsymbol{\theta}^{o}\right)\right\|\leq\left\|\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{j}\left(\tilde{\boldsymbol{\theta}}\right)\right\|+\left\|\boldsymbol{H}_{j}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{j}\left(\boldsymbol{\theta}^{o}\right)\right\|.$$

The desired results then follow from the strong LLN $\sup_{\tilde{\boldsymbol{\theta}}\in\boldsymbol{\Theta}}\left\|\boldsymbol{H}_{Nj,2}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{H}_{j}\left(\tilde{\boldsymbol{\theta}}\right)\right\|\to_{p}\mathbf{0}_{q\times q}$ and $\tilde{\boldsymbol{\theta}}$ is a uniformly consistent estimator of $\boldsymbol{\theta}^{o}$. *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 $\nabla_{\boldsymbol{\theta}} d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)$ in the square brackets,

$$\boldsymbol{q}_{Nj}\left(\boldsymbol{\theta}^o\right) = \frac{1}{N}\sum\nolimits_{n=1}^{N} \tau_j^2\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \hat{d}_{j,-n}\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\left[\nabla_{\boldsymbol{\theta}} \hat{d}_{j,-n}\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right] \tag{S.D.5}$$

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

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

$$\equiv \boldsymbol{q}_{Nj,1}\left(\boldsymbol{\theta}^o\right) + \boldsymbol{q}_{Nj,2}\left(\boldsymbol{\theta}^o\right).$$

We will show that the term $\boldsymbol{q}_{Nj,1}\left(\boldsymbol{\theta}^o\right)$ on the right side of (S.D.5) contributes to the asymptotic distribution while the term $\boldsymbol{q}_{Nj,2}\left(\boldsymbol{\theta}^o\right)$ is asymptotically negligible.

Lemma S.D.6 *Under Assumptions I, TR and E1-E5, we have*

$$\boldsymbol{q}_{Nj,2}\left(\boldsymbol{\theta}^o\right) \equiv \frac{1}{N}\sum\nolimits_{n=1}^{N} \tau_j^2\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \hat{d}_{j,-n}\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)$$
$$\times\left[\nabla_{\boldsymbol{\theta}} \hat{d}_{j,-n}\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \nabla_{\boldsymbol{\theta}} d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right] = o_p\left(N^{-1/2}\right)$$

Proof of Lemma S.D.6: Note that

$$\hat{d}_{j,-n}\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \tag{S.D.6}$$

$$= \hat{d}_{j,-n}\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)$$

$$= \left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right] - \left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right) - \varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\right]$$

$$\equiv \hat{d}_{j,o,-n}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \hat{d}_{j,cs,-n}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right),$$

where the first equality in (S.D.6) holds by $d_j\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) = 0$ 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

$$
\begin{aligned}
&\boldsymbol{q}_{Nj,2}\left(\boldsymbol{\theta}^{o}\right) \qquad \text{(S.D.7)}\\
&=\frac{1}{N}\sum\nolimits_{n=1}^{N}\tau_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right.\\
&\quad\left.+\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)-\varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)\right]\\
&\quad\times\left[\nabla_{\boldsymbol{\theta}}\hat{d}_{j,-n}\left(\boldsymbol{\theta}^{o};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\nabla_{\boldsymbol{\theta}}d_j\left(\boldsymbol{\theta}^{o};\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right]\\
&=A_1+A_2+A_3+A_4
\end{aligned}
$$

where

$$
\begin{aligned}
A_1&=\frac{1}{N}\sum\nolimits_{n=1}^{N}\left[\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\right]\\
&\quad\times\left[\sum\nolimits_{j=1}^{J}\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\nabla_{\boldsymbol{\theta}}\hat{\varphi}_{j,o,-n,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\nabla_{\boldsymbol{\theta}}\varphi_{j,o,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\times\boldsymbol{x}_{n,j}^{(j)}\right]\\
A_2&=\frac{1}{N}\sum\nolimits_{n=1}^{N}\left[\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\right]\\
&\quad\times\left[\sum\nolimits_{j=1}^{J}\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\nabla_{\boldsymbol{\theta}}\hat{\varphi}_{j,cs,-n,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)-\nabla_{\boldsymbol{\theta}}\varphi_{j,cs,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)\right)\times\boldsymbol{x}_{n,j}^{(j)}\right]\\
A_3&=\frac{1}{N}\sum\nolimits_{n=1}^{N}\left[\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)-\varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)\right)\right]\\
&\quad\times\left[\sum\nolimits_{j=1}^{J}\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\nabla_{\boldsymbol{\theta}}\hat{\varphi}_{j,o,-n,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\nabla_{\boldsymbol{\theta}}\varphi_{j,o,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\times\boldsymbol{x}_{n,j}^{(j)}\right]\\
A_4&=\frac{1}{N}\sum\nolimits_{n=1}^{N}\left[\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)-\varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)\right)\right]\\
&\quad\times\left[\sum\nolimits_{j=1}^{J}\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\nabla_{\boldsymbol{\theta}}\hat{\varphi}_{j,cs,-n,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)-\nabla_{\boldsymbol{\theta}}\varphi_{j,cs,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^{o}\right)\right)\times\boldsymbol{x}_{n,j}^{(j)}\right]
\end{aligned}
$$

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

$$\frac{1}{N}\sum_{n=1}^{N}\left[\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\right]^2$$

$$\leq\left(\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|\right)^2=o_p\left(N^{-1/2}\right)$$

and in addition,

$$\frac{1}{N}\sum_{n=1}^{N}\left[\sum_{j=1}^{J}\tau_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\left(\nabla_{\boldsymbol{\theta}}\hat{\varphi}_{j,o,-n,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\nabla_{\boldsymbol{\theta}}\varphi_{j,o,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\times\boldsymbol{x}_{n,1}^{(j)}\right]^2$$

$$\leq\boldsymbol{C}_q\left(\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left\|\nabla_{\boldsymbol{\theta}}\hat{\varphi}_{j,o,-n,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\nabla_{\boldsymbol{\theta}}\varphi_{j,o,(j)}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right\|\right)^2=o_p\left(N^{-1/2}\right)$$

where $\boldsymbol{C}_q\in\mathbb{R}^q$. 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 $\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)=\tau_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\nabla_{\boldsymbol{\theta}}d_j\left(\boldsymbol{\theta}^o;\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)$ for notational simplicity. By (S.D.6) we have

$$\boldsymbol{q}_{Nj,1}\left(\boldsymbol{\theta}^o\right)=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\hat{d}_{j,-n}\left(\boldsymbol{\theta}^o;\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\tag{S.D.8}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\right].$$

Lemma S.D.7 *Under Assumptions I, TR and E1-E5,*

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

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

$$\psi_{N,o}(\omega_m,\omega_n) = \boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\left(y_{mj} - \varphi_{j,o}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right)$$

$$\times K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)} - \boldsymbol{z}_n^{(j)}\right) K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)} - \boldsymbol{X}_n^{(j)}\right) f_j^{-1}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right),$$

$$\psi_{N,cs}(\omega_m,\omega_n) = \boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\left(y_{mj} - \varphi_{j,cs}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\right)$$

$$\times K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)} - \left(-\boldsymbol{z}_n^{(j)} - 2\boldsymbol{X}_n^{(j)}\boldsymbol{\theta}^o\right)\right) K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)} - \boldsymbol{X}_n^{(j)}\right) f_j^{-1}\left(-\boldsymbol{z}_n^{(j)} - 2\boldsymbol{X}_n^{(j)}\boldsymbol{\theta}^o, \boldsymbol{X}_n^{(j)}\right).$$

where $K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)} - \cdot\right) = \prod_{l=1}^J h_N^{-2J} k^{(1)}\left(h_{z_l}^{-1}\left(z_{ml}^{(j)} - \cdot\right)\right)$ *where* $k^{(1)}$ *is the first derivative of kernel function.*

Proof of Lemma S.D.7: We first observe that

$$\boldsymbol{q}_{Nj,1}(\boldsymbol{\theta}^o) = \frac{1}{N}\sum_{n=1}^N \boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right) \hat{d}_{j,-n}\left(\boldsymbol{\theta}^o; \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) \tag{S.D.9}$$

$$= \frac{1}{N}\sum_{n=1}^N \boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\right].$$

$$= \frac{1}{N}\sum_{n=1}^N \boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\left\{\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right]\right.$$

$$\left. - \left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right) - \varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\right]\right\}$$

$$= \frac{1}{N}\sum_{n=1}^N \boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\left\{\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \mathbb{E}\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\middle| \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right]\right.\right.$$

$$\left. + \mathbb{E}\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\middle| \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right] - \varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\right]$$

$$- \left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right) - \mathbb{E}\left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\middle| \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right]\right.$$

$$\left.\left. + \mathbb{E}\left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\middle| \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right] - \varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\right]\right\}$$

$$= \frac{1}{N}\sum_{n=1}^N \boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\left\{\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right) - \mathbb{E}\left[\hat{\varphi}_{j,o,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right)\middle| \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right]\right]\right.$$

$$\left. - \left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right) - \mathbb{E}\left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}, \boldsymbol{\theta}^o\right)\middle| \boldsymbol{z}_n^{(j)}, \boldsymbol{X}_n^{(j)}\right]\right]\right\} + O(h^s)$$

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

$$\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\mathbb{E}\left[\hat{\varphi}^{(J)}_{j,o,-n}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\middle|\,\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]-\varphi^{(J)}_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|=O\left(h^s\right)$$

$$\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\mathbb{E}\left[\hat{\varphi}^{(J)}_{j,cs,-n}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\middle|\,\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]-\varphi^{(J)}_{j,cs}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\right|=O\left(h^s\right)$$

Next, to derive $\hat{\varphi}^{(J)}_{j,o,-n}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-\mathbb{E}\left[\hat{\varphi}^{(J)}_{j,o,-n}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\middle|\,\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]$, we observe that

$$\hat{\varphi}^{(J)}_{j,o,-n}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)-\mathbb{E}\left[\hat{\varphi}^{(J)}_{j,o,-n}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\middle|\,\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]$$

$$=\hat{f}_j^{-1}\frac{1}{N}\sum\nolimits_{m=1,m\neq n}^{N}y_{mj}K^{(J)}_{\boldsymbol{h_z}}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)$$

$$-E\left[\hat{\varphi}^{(J)}_{j,o,-n}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)|\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]$$

$$=f_j^{-1}\frac{1}{N}\sum\nolimits_{m=1,m\neq n}^{N}\left(y_{mj}-\varphi^{(J)}_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)$$

$$\times K^{(J)}_{\boldsymbol{h_z}}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)+R_{o,1}$$

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 $\hat{f}_j^{-1}$.[1] Note that $R_{o,1}$ is bounded by the product of

$$\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\mathbb{E}\left[\hat{\varphi}^{(J)}_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\middle|\,\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]-\varphi^{(J)}_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|,$$

and

$$\sup_{\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\in\mathcal{S}^{Tr}_{\left(\boldsymbol{z}^{(j)},\boldsymbol{X}^{(j)}\right)}}\left|\hat{f}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-f_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|.$$

[1] $\hat{f}_j^{-1}=f_j^{-1}\left(1-f_j^{-1}\left(\hat{f}_j-f_j\right)+2f_j^{-2}\left(\hat{f}_j-f_j\right)^2+o\left(\hat{f}_j-f_j\right)^2\right)$

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

$$\psi_{N,o,1}\left(\omega_m,\omega_n\right)=\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)$$

$$\times f_j^{-1}\left(y_{mj}-\varphi_{j,o}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)K_{\boldsymbol{h}_{\boldsymbol{z}}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)K_{\boldsymbol{h}_{\boldsymbol{X}}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)$$

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

In addition, to derive $\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)-E\left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\middle|\,\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]$, we observe that

$$\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)-\mathbb{E}\left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\middle|\,\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]$$

$$=\hat{f}_j^{-1}y_{mj}K_{\boldsymbol{h}_{\boldsymbol{z}}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\left(-\boldsymbol{z}_n^{(j)}-2\boldsymbol{\theta}^o\boldsymbol{X}_n^{(j)}\right)\right)K_{\boldsymbol{h}_{\boldsymbol{X}}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)$$

$$-\mathbb{E}\left[\hat{\varphi}_{j,cs,-n}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)|\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right]$$

$$=f_j^{-1}\frac{1}{N}\sum\nolimits_{m=1,m\neq n}^{N}\left(y_{mj}-\varphi_{j,cs}^{(J)}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\right)$$

$$\times K_{\boldsymbol{h}_{\boldsymbol{z}}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\left(-\boldsymbol{z}_n^{(j)}-2\boldsymbol{\theta}^o\boldsymbol{X}_n^{(j)}\right)\right)K_{\boldsymbol{h}_{\boldsymbol{X}}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)+R_{cs,1}$$

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 decompostion of $\hat{f}_j^{-1}$, with the order of $O_p\left(N^{-1/2}\right)$, by the same argument as above. Denoting

$$\psi_{N,cs,1}\left(\omega_m,\omega_n\right)=\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)$$

$$\times f_j^{-1}\left(y_{mj}-\varphi_{j,cs}^{(J)}\right)K_{\boldsymbol{h}_{\boldsymbol{z}}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\left(-\boldsymbol{z}_n^{(j)}-2\boldsymbol{\theta}^o\boldsymbol{X}_n^{(j)}\right)\right)K_{\boldsymbol{h}_{\boldsymbol{X}}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)$$

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.D.8 *Under Assumptions E2-E5,*

$$\frac{1}{N\left(N-1\right)}\sum\nolimits_{m=1}^{N}\sum\nolimits_{n=1,n\neq m}^{N}\psi_N\left(\omega_m,\omega_n\right)=\frac{1}{N}\sum\nolimits_{m=1}^{N}\boldsymbol{t}_{mj}+o_p\left(N^{-1/2}\right)$$

and

$$N^{-1/2}\sum\nolimits_{m=1}^{N}\boldsymbol{t}_{mj}\rightarrow_d N\left(\mathbf{0}_q,\boldsymbol{\Omega}_j\right),$$

where $\boldsymbol{t}_{mj}=\left(\upsilon_{mj,o}-\upsilon_{mj,cs}\right)\partial^J\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)/\partial z_1^{(j)}\cdots\partial z_J^{(j)}$ *and* $\boldsymbol{\Omega}_j=\mathbb{E}\left[\boldsymbol{t}_{mj}\boldsymbol{t}_{mj}'\right]$.

Proof of Lemma S.D.8: We denote $U_n\left(\omega_m,\omega_n\right)$ as the second-order U-statistic and $\hat{U}_n\left(\omega_m,\omega_n\right)$ as the projection of the second-order U-statistic, which is given by

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

and

$$\hat{U}_n=\mathbb{E}\left[\psi_N\left(\omega_m,\omega_n\right)\right]+\frac{1}{N}\sum\nolimits_{m=1}^{N}\left(r_{N1}\left(\omega_m\right)-\mathbb{E}\left[\psi_N\left(\omega_m,\omega_n\right)\right]\right)$$
$$+\frac{1}{N}\sum\nolimits_{n=1}^{N}\left(r_{N2}\left(\omega_n\right)-\mathbb{E}\left[\psi_N\left(\omega_m,\omega_n\right)\right]\right),$$

where $r_{N1}\left(\omega_m\right)=\mathbb{E}\left[\psi_N\left(\omega_m,\omega_n\right)|\omega_m\right]$ and $r_{N2}\left(\omega_n\right)=\mathbb{E}\left[\psi_N\left(\omega_m,\omega_n\right)|\omega_n\right]$. To apply this to Lemma S.D.1, we first show that $\mathbb{E}\left[\left\|\psi_N\left(\omega_m,\omega_n\right)\right\|^2\right]=o(N)$, which is equivalent to showing that $\mathbb{E}\left[\left\|\psi_{N,o}\left(\omega_m,\omega_n\right)\right\|^2\right]=o(N)$ and $\mathbb{E}\left[\left\|\psi_{N,cs}\left(\omega_m,\omega_n\right)\right\|^2\right]=o(N)$. Recall that $\boldsymbol{h}_{\boldsymbol{z}}=\left(h_N,\cdots,h_N\right)'$ and $\boldsymbol{h}_{\boldsymbol{X}}=\left(h_N,\cdots,h_N,\cdots,h_N,\cdots,h_N\right)'$ in Assumption E4. Denote $\boldsymbol{u}_{\boldsymbol{z}}^{(j)}=h_N^{-1}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)$ and $\boldsymbol{u}_{\boldsymbol{X}}^{(j)}=h_N^{-1}\left[\wedge\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)\right]$, where $\wedge$ is juxtaposing the consecutive rows of the matrix next to each other. In addition, define $\bar{\wedge}$ the inverse transformation (stacking the vector into a matrix) of $\wedge$. By direct calculation

$$\begin{aligned}
&\mathbb{E}\left[\left\|\psi_{N,o}\left(\omega_m,\omega_n\right)\right\|^2\right] \qquad \text{(S.D.10)}\\
&=\int\left\|K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)\right\|^2\\
&\times\left[\varphi_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)+\varphi_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)-2\varphi_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)\varphi_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right]\\
&\times f_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)f_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)f_j^{-1}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\boldsymbol{\xi}_j^2\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)d\boldsymbol{z}_m^{(j)}d\boldsymbol{X}_m^{(j)}d\boldsymbol{z}_n^{(j)}d\boldsymbol{X}_n^{(j)}\\
&\quad=\int\left\|K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{u_z}^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{u_X}^{(j)}\right)\right\|^2\\
&\quad\times\left[\varphi_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)+\varphi_j^2\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{u_z}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_m^{(j)}-\bar{\wedge}\left(\boldsymbol{u_X}^{(j)}\boldsymbol{h_X}\right)\right)\right.\\
&\quad\left.-2\varphi_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)\varphi_j\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{u_z}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_m^{(j)}-\bar{\wedge}\left(\boldsymbol{u_X}^{(j)}\boldsymbol{h_X}\right)\right)\right]\\
&\quad\times f_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)\boldsymbol{\xi}_j^2\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{u_z}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_m^{(j)}-\bar{\wedge}\left(\boldsymbol{u_X}^{(j)}\boldsymbol{h_X}\right),\boldsymbol{\theta}^o\right)d\boldsymbol{z}_m^{(j)}d\boldsymbol{X}_m^{(j)}d\boldsymbol{u_z}^{(j)}d\boldsymbol{u_X}^{(j)}\\
&\quad=O\left(h_N^{-2J-J-Jq}\right)=O\left(N\left(Nh_N^{2J+J+rfJq}\right)^{-1}\right)=o\left(N\right),
\end{aligned}$$

where the first equality in (S.D.10) follows from definitions; the second equality holds using a change of variables; and the third equality is satisfied by Assumptions E3 and E4. The desired result then follows from Assumption E4. Similarly, we can show $\mathbb{E}\left[\left\|\psi_{N,cs}\left(\omega_m,\omega_n\right)\right\|^2\right]=o(N)$.

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) $\mathbb{E}\left[\psi_N\left(\omega_m,\omega_n\right)\right]=\mathbb{E}\left[r_{N1}\left(\omega_m\right)\right]=\mathbb{E}\left[r_{N2}\left(\omega_n\right)\right]=o\left(N^{-1/2}\right)$, (ii) $\frac{1}{N}\sum_{n=1}^N\left(r_{N2}\left(\omega_n\right)-\mathbb{E}\left[r_{N2}\left(\omega_n\right)\right]\right)=o_p\left(N^{-1/2}\right)$, and (iii) $\frac{1}{N}\sum_{m=1}^N\left(r_{N1}\left(\omega_m\right)-E\left[r_{N1}\left(\omega_m\right)\right]\right)=N^{-1}\sum_{m=1}^N\boldsymbol{t}_{mj}$,where $N^{-1/2}\sum_{m=1}^N\boldsymbol{t}_{mj}$ is $N\left(\mathbf{0}_q,\boldsymbol{\Omega}_j\right)$.

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

$$
\begin{aligned}
\mathbb{E}\left[\psi_{N,o}\left(\omega_m,\omega_n\right)\right] = & \ \mathbb{E}\left[\mathbb{E}\left[\psi_{N,o}\left(\omega_m,\omega_n\right)|\omega_n\right]\right] \\
& = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\left(y_{mj}-\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\right.\right. \\
& \left.\left.\times K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)\times f_j^{-1}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|\omega_n\right]\right] \\
& = \mathbb{E}\left[\int\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\left(\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)}+\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_n^{(j)}+\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right)\right)-\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\right. \\
& \times K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right)f_j^{-1}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right) \\
& \left.\times\ f_j^{-1}\left(\boldsymbol{z}_n^{(j)}+\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_n^{(j)}+\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right)\right)d\boldsymbol{u}_{\boldsymbol{z}}^{(j)}d\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right|\omega_n\right] \\
& = -\mathbb{E}\left[\int\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\frac{\partial\left(\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)}+\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_n^{(j)}+\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right)\right)-\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)}{\partial\boldsymbol{u}_{\boldsymbol{z}}^{(j)}}\right. \\
& \left.\times\ K_{\boldsymbol{h_z}}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right)f_j^{-1}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right|\omega_n\right] \\
& = O\left(h_N^{\boldsymbol{s}}\right).
\end{aligned}
$$

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

$$
\mathbb{E}\left[\psi_N\left(\omega_m,\omega_n\right)\right] = \mathbb{E}\left[\psi_{N,o}\left(\omega_m,\omega_n\right)\right] - \mathbb{E}\left[\psi_{N,cs}\left(\omega_m,\omega_n\right)\right] = O\left(h_N^{\boldsymbol{s}}\right).
$$

Second, to show Part (ii) holds, by direct calculation, we have

$$
\begin{aligned}
r_{N2,o}(\omega_n) = {} & \mathbb{E}\left[\psi_{N,o}(\omega_m,\omega_n)|\omega_n\right] \qquad \text{(S.D.11)}\\
& = \mathbb{E}\Big[\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\left(y_{mj}-\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\\
& \times K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)f_j^{-1}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\Big|\,\omega_n\Big]\\
& = \mathbb{E}\Big[\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\left(\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)-\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\\
& \times K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)f_j^{-1}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\Big|\,\omega_n\Big]\\
& = \int \boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\frac{\partial\left(\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)}+\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_n^{(j)}+\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right)\right)-\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)}{\partial\boldsymbol{u}_{\boldsymbol{z}}^{(j)}}\\
& \times K_{\boldsymbol{h_z}}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right)f_j^{-1}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\\
& \times f_j\left(\boldsymbol{z}_n^{(j)}+\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_n^{(j)}+\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right)\right)d\boldsymbol{u}_{\boldsymbol{z}}^{(j)}d\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\\
& = O\left(h_N^{\boldsymbol{s}}\right) = o\left(N^{-1/2}\right).
\end{aligned}
$$

The last second equality follows from integration by parts and a Taylor expansion. We therefore get that $\frac{1}{N}\sum_{n=1}^{N}\left(r_{N2}(\omega_n)-\mathbb{E}\left[r_{N2}(\omega_n)\right]\right)=o_p\left(N^{-1/2}\right)$.

To show Part (iii), we have

$$
\begin{aligned}
r_{N1,o}(\omega_m) &= \mathbb{E}\left[\psi_{N,o}(\omega_m,\omega_n)\,|\omega_m\right] \qquad \text{(S.D.12)}\\
&= \mathbb{E}\Big[\boldsymbol{\xi}_j\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)},\boldsymbol{\theta}^o\right)\left(y_{mj}-\varphi_{j,o}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\right)\\
&\times K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{z}_n^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{X}_m^{(j)}-\boldsymbol{X}_n^{(j)}\right)f_j^{-1}\left(\boldsymbol{z}_n^{(j)},\boldsymbol{X}_n^{(j)}\right)\Big|\,\omega_m\Big]\\
&= \int \boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_m^{(j)}-\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right),\boldsymbol{\theta}^o\right)\\
&\times\left(y_{mj}-\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_m^{(j)}-\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right)\right)\right)\\
&\times K_{\boldsymbol{h_z}}^{(J)}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right)d\boldsymbol{u}_{\boldsymbol{z}}d\boldsymbol{u}_X\\
&= \int\left[\frac{\partial^J\left(\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_m^{(j)}-\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right),\boldsymbol{\theta}^o\right)\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_m^{(j)}-\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right)\right)\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right]\\
&\times K_{\boldsymbol{h_z}}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right)d\boldsymbol{u}_{\boldsymbol{z}}^{(j)}d\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\\
&- y_{mj}\int\frac{\partial^J\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}-\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h_z},\boldsymbol{X}_m^{(j)}-\bar{\wedge}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h_X}\right),\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}K_{\boldsymbol{h_z}}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right)K_{\boldsymbol{h_X}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right)d\boldsymbol{u}_{\boldsymbol{z}}^{(j)}d\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\\
&= r_o(\omega_m)+\varsigma_{N,o}(\omega_m)
\end{aligned}
$$

where

$$r_o(\omega_m) = \int \frac{\partial^J \boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}, \boldsymbol{\theta}^o\right)\varphi_j\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}} K_{\boldsymbol{h}_{\boldsymbol{z}}}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right) K_{\boldsymbol{h}_{\boldsymbol{X}}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right) d\boldsymbol{u}_{\boldsymbol{z}}^{(j)} d\boldsymbol{u}_{\boldsymbol{X}}^{(j)} \tag{S.D.13}$$

$$- y_{mj}\int \frac{\partial^J \boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}, \boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}} K_{\boldsymbol{h}_{\boldsymbol{z}}}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right) K_{\boldsymbol{h}_{\boldsymbol{X}}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right) d\boldsymbol{u}_{\boldsymbol{z}}^{(j)} d\boldsymbol{u}_{\boldsymbol{X}}^{(j)}$$

$$= \boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}, \boldsymbol{\theta}^o\right)\frac{\partial^J \varphi_{j,o}\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}$$

$$- \left(y_{mj} - \varphi_{j,o}\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}\right)\right)\frac{\partial^J \boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}, \boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}$$

and

$$\varsigma_{N,o}(\omega_m) = \int \frac{\partial^J\left[\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)} - \boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{X}_m^{(j)} - \barwedge\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h}_{\boldsymbol{X}}\right), \boldsymbol{\theta}^o\right)\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)} - \boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{X}_m^{(j)} - \barwedge\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h}_{\boldsymbol{X}}\right)\right)\right]}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}} \tag{S.D.14}$$

$$\left.-\frac{\partial^J\left[\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}, \boldsymbol{\theta}^o\right)\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}\right)\right]}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right] K_{\boldsymbol{h}_{\boldsymbol{z}}}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right) K_{\boldsymbol{h}_{\boldsymbol{X}}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right) d\boldsymbol{u}_{\boldsymbol{z}}^{(j)} d\boldsymbol{u}_{\boldsymbol{X}}^{(j)}$$

$$- y_{m,j}\int\left[\frac{\partial^J \boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)} - \boldsymbol{u}_{\boldsymbol{z}}^{(j)}\boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{X}_m^{(j)} - \barwedge\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\boldsymbol{h}_{\boldsymbol{X}}\right), \boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}} - \frac{\partial^J \boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)}, \boldsymbol{X}_m^{(j)}, \boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right]$$

$$\times K_{\boldsymbol{h}_{\boldsymbol{z}}}\left(\boldsymbol{u}_{\boldsymbol{z}}^{(j)}\right) K_{\boldsymbol{h}_{\boldsymbol{X}}}\left(\boldsymbol{u}_{\boldsymbol{X}}^{(j)}\right) d\boldsymbol{u}_{\boldsymbol{z}}^{(j)} d\boldsymbol{u}_{\boldsymbol{X}}^{(j)}$$

Then it follows that

$$\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(r_{N1,o}(\omega_m) - \mathbb{E}\left[r_{N1,o}(\omega_m)\right]\right) = \frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(r_o(\omega_m) - \mathbb{E}\left[r_o(\omega_m)\right]\right)$$

$$+ \frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(\varsigma_{N,o}(\omega_m) - \mathbb{E}\left[\varsigma_{N,o}(\omega_m)\right]\right).$$

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

$$\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(r_o\left(\omega_m\right)-\mathbb{E}\left[r_o\left(\omega_m\right)\right]\right) \tag{S.D.15}$$

$$=\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)\frac{\partial^J\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right.$$

$$\left.-\mathbb{E}\left[\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)\frac{\partial^J\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right]\right)$$

$$-\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(y_{mj}-\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)\right)\frac{\partial^J\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}$$

$$+\mathbb{E}\left[\left(y_{mj}-\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)\right)\frac{\partial^J\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right]$$

and

$$\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(r_{cs}\left(\omega_m\right)-\mathbb{E}\left[r_{cs}\left(\omega_m\right)\right]\right) \tag{S.D.16}$$

$$=\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)\frac{\partial^J\varphi_{j,cs}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right.$$

$$\left.-\mathbb{E}\left[\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)\frac{\partial^J\varphi_{j,cs,}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right]\right)$$

$$-\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(y_{mj}-\varphi_{j,cs}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)\right)\frac{\partial^J\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}$$

$$+\mathbb{E}\left[\left(y_{mj}-\varphi_{j,cs}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)\right)\frac{\partial^J\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}\right]$$

Then

$$\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(r\left(\omega_m\right)-\mathbb{E}\left[r\left(\omega_m\right)\right]\right)$$

$$=\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(r_o\left(\omega_m\right)-\mathbb{E}\left[r_o\left(\omega_m\right)\right]\right)-\frac{1}{\sqrt{N}}\sum\nolimits_{i=1}^{N}\left(r_{cs}\left(\omega_m\right)-\mathbb{E}\left[r_{cs}\left(\omega_m\right)\right]\right)$$

$$=-\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(y_{mj}-\varphi_{j,o}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)}\right)\right)\frac{\partial^J\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}$$

$$+\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\left(y_{mj}-\varphi_{j,cs}\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)\right)\frac{\partial^J\boldsymbol{\xi}_j\left(\boldsymbol{z}_m^{(j)},\boldsymbol{X}_m^{(j)},\boldsymbol{\theta}^o\right)}{\partial z_1^{(j)}\cdots\partial z_J^{(j)}}$$

$$=\frac{1}{\sqrt{N}}\sum\nolimits_{m=1}^{N}\boldsymbol{t}_{mj}.$$

The first two terms in equation (S.D.15) cancel out with the first two terms in equation (S.D.16) by the identification equation. In addition, by Assumption E5 (similar to Assumption 3 in Powell et al. (1989)), this last term $\frac{1}{\sqrt{N}}\sum_{m=1}^{N}\left(\varsigma_{N,o}\left(\omega_m\right)-\mathbb{E}\left[\varsigma_{N,o}\left(\omega_m\right)\right]\right)$ has second moment, that is bounded by $4h_z^{2J+2Jq}\left\{\mathbb{E}\left[\left(1+|y|+\|\boldsymbol{z}\|\right)m\left(\boldsymbol{z},\boldsymbol{\cdot}\right)\right]^2\left[\int\|\boldsymbol{u}\|\,|K\left(\boldsymbol{u}\right)|d\boldsymbol{u}\right]^2\right\}=O\left(h_z^{2J+2Jq}\right)$. So it will converge to zero in probability. Applying Linderberg-Feller Central Limit Theorem using Assumption E2 gives the desired results, where $\boldsymbol{\Omega}_j=\mathbb{E}\left[\boldsymbol{t}_{mj}\boldsymbol{t}_{mj}'\right]$. *Q.E.D.*

Proof of Theorem S.B.3: The limit distribution in Theorem S.B.3 then follows from Lemmas S.D.5–S.D.8 and the non-singularity of $\boldsymbol{H}_j$ in Assumption E6 as well as the symmetry of the indices m and n. *Q.E.D.*

Additional References

Andrews, D.W.K. (1994): "Empirical Process Methods in Economics," *Handbook of Econometrics*, Vol. 4. ed. by R.F.Engle and D.L.McFadden. Elsevier Science B. V. 2247-2294.

Chen, S., Khan, S., and Tang, X. (2016): "Informational Content of Special Regressors in Heteroskedastic Binary Response Models," *Journal of Econometrics*, 193, 162-182.

Hong, H. and Tamer E. (2003): "Inference in Censored Models with Endogenous Regressors," *Econometrica*, 71(3), 905-932.

Powell, J.L., Stock, J., and Stocker, T. (1989): "Semiparametric Estimation of Index Models," *Econometrica*, 57, 1403-1430.

Newey, W.K. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10(2), 233-253

Newey, W.K. and McFadden, D. (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, Vol. 4, 2111-2245.