WARWICK
ECONOMICS

Fixed Effects Nonlinear Panel Models with Heterogeneous Slopes: Identification and Consistency

Martin Mugnier & Ao Wang

December 2024 No: 1531

Warwick Economics Research Papers

ISSN 2059-4283 (online)

ISSN 0083-7350 (print)

Fixed Effects Nonlinear Panel Models with Heterogeneous Slopes: Identification and Consistency*

Martin Mugnier† Ao Wang‡

Abstract

We study a class of two-way fixed effects index function models with a nonparametric link function and individual- (or time-) specific slopes. Our model alleviates potential misspecification errors due to the common practice of specifying a known link function such as Gaussian and its tail behavior. It also enables to incorporate richer unobserved heterogeneity in the marginal effects of covariates via heterogeneous slopes across individuals. We show the identification of the link function as well as the slopes and fixed effects parameters when both individual and time dimensions are large. We propose a nonparametric consistency result for the fixed effects sieve maximum likelihood estimators. Finally, we apply our method to the study of establishing exportation and illustrate the consequences of imposing Gaussian link function and homogeneity on the slope of distance.

Keywords: Nonlinear Panel Models, Fixed Effects, Slope Heterogeneity, Nonparametric, Sieve.

JEL Codes: C23, C24, C25.

*This paper was previously circulated as *Identification and (Fast) Estimation of Large Nonlinear Panel Models with Two-Way Fixed Effects*.

†Paris School of Economics, martin.mugnier@psemail.eu

‡Corresponding author; University of Warwick and CAGE Research Centre, ao.wang@warwick.ac.uk

1 Introduction

Nonlinear two-way fixed effects panel models gain popularity in economic research. These models typically feature individual and time dimensions, enabling researchers to incorporate rich heterogeneity in empirical research, e.g., industrial organization (Dubois et al., 2020), international trade (Helpman et al., 2008), labor (Abowd et al., 1999), innovation (Aghion et al., 2013) and network (Jochmans, 2018). By allowing both dimensions to increase to infinity, one can reduce the incidental parameter problem in panel data models (Lancaster, 2000; Neyman and Scott, 1948) to a post-estimation bias correction (Fernández-Val and Weidner, 2016).

In applied and related econometric works in this literature, researchers often adopt two restrictions: parametrically specified link function (e.g., probit) and slope homogeneity across individuals and time (e.g., homogeneous price coefficient in demand). These restrictions may not be innocuous. For instance, in the setting of demand, if the true link function has a relatively thick left tail, assuming a thin-tail link function (e.g., Gaussian) may understate the negative effect of price and introduce an upward bias in the estimated price coefficient. Imposing the same price coefficient across individuals may overlook unobserved heterogeneity in price sensitivity and lead to biased estimates of average marginal effects of price. The extent to which these restrictions can be relaxed is, however, still underexplored in the literature.

In this paper, we study the nonparametric identification and estimation of a class of two-way fixed effects index function models that relax the aforementioned two restrictions. In this class of models, individual i's probability of choosing y from a discrete set $\mathcal{Y}$ at time t is given by:

$$\Pr\left(Y_{it} = y | (X_{is})_{s \leq t}, \mathcal{F}\right) = g(y; X_{it}'\beta_i + \alpha_i + \xi_t), \tag{1}$$

where X_{it} are individual i's observed characteristics at time t, β_i are individual-specific slopes, (α_i, ξ_t) are individual-fixed and time-fixed effects, g is a link function, and $\mathcal{F}$ is the smallest $\sigma-$field generated by latent shocks common to all individuals (e.g., $\{\xi_t\}_t$) and those common to all time series (e.g., $\{\alpha_i\}_i$). This model encompasses settings with a single index, such as binary outcome, ordered outcome and count outcome, as well as those with multinomial outcomes. Contrasting the common practice of specifying a known link function, model (1) allows

g to be nonparametric and estimated from the data. Besides, we allow slope parameters β_i to be individual-specific rather than homogeneous across individuals (i.e., $\beta_i = \beta$).[1] This feature enables researchers to incorporate richer (unobserved) individual heterogeneity in the marginal effects of covariates of interest, e.g., household's price sensitivity and trade cost.

First, we lay out the asymptotic framework within which the identification of $(\beta_i, \alpha_i)_i$, $(\xi_t)_t$, and $g(\cdot;\cdot)$ in (1) is meaningful when both the numbers of individuals (N) and time periods (T) are large. This framework is characterized by a well-defined infinite population (Assumption 1) and sequences of increasing consistent samples (Assumption 2). It allows to construct individual-specific limiting objects, e.g., the individual-specific probability of observing outcome $y_{it} = y$ given a vector of covariate values, from the individual-specific time series. Similarly, one can also construct time-specific limiting objects from the cross-sectional observations in each time period. Our identification arguments will rely on these limiting objects. Such constructions do not require the knowledge of the true parameters. Moreover, we do not impose stationary distribution over the time dimension and allow for lagged outcomes as explanatory covariates.

Second, we propose two identification results (Theorems 1 and 2) for $(\alpha_i, \beta_i)_i$, $(\xi_t)_t$, and function g when both N and T are infinity. In Theorem 1, we adopt the technique of *compensating variable* to establish the identification of (α_i, β_i) relative to some individual j's (α_j, β_j). Loosely speaking, we require the existence of a variable in X_{jt}, say its first component $X_{jt}^{(1)}$, that can compensate the difference in i's and j's indexes due to (α_i, β_i)–(α_j, β_j). This compensating variable can be a continuous one (e.g., the distance in trade) and does not need to have a large support. Under some monotonicity assumptions on the link function with respect to the index, one can back out the amount of the compensation by comparing identified i's and j's limiting objects, giving rise to restrictions on (α_i, β_i) and identifying their values relative to (α_j, β_j). More generally, this pairwise argument induces a compensating network in which two individuals are connected if and only if one can compensate the other. The ability of relatively identifying individual-specific parameters is then translated to the connectedness of the network. Within its connecting component, one can achieve the identification of (α_i, β_i) relative to the parameters of a reference individual in this component. The compensating network may have more than one disjoint connecting components. In this case, the

[1] In Appendix E, we also consider the case of time-specific slopes.

corresponding reference individuals' parameters are not identified without further conditions.

In Theorem 2, we assume the connectedness of the compensating network and propose conditions to further identify time-specific parameters $(\xi_t)_t$ and the link function g. In a similar vein to Theorem 1 and given the identification of individual-specific parameters, we apply the argument of compensating variable to achieve the identification of $(\xi_t)_t$ relative to a reference time period. We then identify the normalized index for each cell (i,t) (relative to the reference individual and time period) and $g(y;v)$ by the probability of observing outcome y given index value v.

Third, we build on the identification results and propose a nonparametric consistency result for the fixed effects sieve maximum likelihood estimators of $(\alpha_i,\beta_i)_i$, $(\xi_t)_t$, and link function g (Theorem 3). The consistency requires N and T to increase to infinity and does not depend on their relative rate. Our result implies the uniform convergences of the maximum likelihood estimators of normalized $(\alpha_i,\beta_i)_i$, $(\xi_t)_t$, and the sieve estimator of g. Consequently, we obtain consistent plug-in estimators for individual-level marginal effects of covariates when both N and T tend to infinity.

Finally, we revisit the study of establishing between-country exportation in Helpman et al. (2008) to illustrate the consequences of imposing a known link function and slope homogeneity. The original paper uses a two-way fixed effects probit model with a constant coefficient on the log-distance between countries. Differently, we allow the link function to be unknown and the distance coefficient to be exporting-country-specific. The estimated link function has a thicker left tail than Gaussian distribution. As a result, imposing the probit specification leads to a upward bias in the estimates of distance coefficients. Besides, we find non-trivial variation in the country-specific marginal effect of distance when allowing for country-specific distance slope. Whether or not we treat the link function as parameters to be estimated, the slope heterogeneity explains more than 90% of this variance. In contrast, the homogeneous probit model substantially underestimates the variation of country-specific marginal effect of distance by a factor of 20 compared to models with country-specific distance slope.

Related literature. Our paper belongs to the strand of research on panel data methods in the large-N and large-T asymptotic framework. In a parametric set-

ting with only individual fixed effects, existing works establish that one can reduce the Neyman and Scott (1948)'s incidental parameters problem to a post-estimation bias correction by allowing T to increase to infinity (Dhaene and Jochmans, 2015; Fernández-Val, 2009; Fernández-Val and Lee, 2013; Hahn and Kuersteiner, 2002; Hahn and Newey, 2004). Recent works extend this result to two-way fixed effects models with known link function (Fernández-Val and Weidner, 2016, 2018), interactive fixed effects (Bai, 2009; Boneva and Linton, 2017; Chen, 2016; Chen et al., 2021; Gao et al., 2023), and the dyadic network formation (e.g., Graham (2017), Jochmans (2018), Zeleneev (2020). See de Paula (2020) for a review).

Different from most aforementioned papers, we consider a two-way fixed effects panel model with a nonparametric link function. We provide conditions for the identification of link function as well as slopes and fixed effects when both N and T are large. Our identification arguments (compensating variable) differ from those in dyadic network formation that rely on specific network features such as undirectedness, i.e., the "time" and individual dimensions coincide (Candelaria, 2020; Gao, 2020; Toth, 2017; Zeleneev, 2020). Building on the identification, we further propose a consistency result for the sieve maximum likelihood estimator. Both results are novel in this literature to the best of our knowledge; they provide a foundation for the parametric large-N-and-large-T inference methods that usually rely on specific properties of the link function such as logit (Charbonneau, 2017; Jochmans, 2018) and log-concavity (e.g., Assumption 4.1 in Fernández-Val and Weidner (2016)).

Besides, unlike most existing two-way fixed effects panel models, ours allow for unobserved heterogeneity in slopes.[2] In this regard, our paper speaks to the literature of random coefficients panel models that aim to incorporate flexible partial effects of covariates across units. This literature usually studies the identification and estimation of distributional features of the random coefficients in parametric one-way fixed effect models mostly in a fixed-T setting (Arellano and Bonhomme, 2011; Chamberlain, 1982; Hsiao and Pesaran, 2004), with few exceptions also considering a large-T asymptotics (Fernández-Val and Lee, 2013; Swamy, 1970). Our paper differs from these works by considering a nonparametric two-way fixed effects model and prove the point identification of heterogeneous slopes when both N and T are large. This result allows to identify and consistently estimate unit-specific

[2]Boneva and Linton (2017) and Gao et al. (2023) also consider such heterogeneity but assume a known link function.

partial effects of covariates, contrasting existing results in the fixed-T setting that only the distributional features of the partial effects are identified (Graham and Powell, 2012). However, we do not tackle the inference about the partial effects in the presence of nonparametric link function and heterogeneous slopes.

Our identification strategy of compensating variable relates to several works beyond panel models. Examples include the identification of the average effects of endogenous dummy covariates (Vytlacil and Yildiz, 2007), testing the exclusion restriction in the control function approach (D'Haultfoeuille et al., 2021), and demand identification in the presence of railway dynamic pricing (D'Haultfœuille et al., 2022). Analogously to these works, our strategy relies on the existence of an exogenous covariate with sufficient variation to compensate changes in other covariates or in fixed effects. It does not require this covariate to have a large support. Besides, the compensating variable resembles conceptually a special regressor (Lewbel, 2014; Williams, 2020). Due to the unobserved heterogeneity in slopes, the slope of the compensating variable may, however, differ across individuals and cannot be normalized to one as in the method of special regressor. Using the cross-sectional variation in the compensating variable, we also compensate the difference in its slopes across individuals, a new aspect the special regressor method (and previously mentioned works) does not have.

2 Model and Sampling Process

Consider a countably infinite population $\{(i,t) : (i,t) \in \{1,2,\ldots\}^2\}$. Each cell (i,t) is equipped with an observed vector $(Y_{it}, X_{it}') \in \mathcal{Y} \times \mathcal{X}$ that is typically random. We assume that there exist σ-fields $\mathcal{F}^{cs}$ and $\mathcal{F}^{ts}$ generated by latent shocks common to all cross-sectional units (the i's) and all time series (the t's) respectively. Let $\mathcal{F}$ be the smallest σ-field containing $\mathcal{F}^{cs} \cup \mathcal{F}^{ts}$.

Assumption 1 (Model)**.**

(a) Single index and two-way fixed effects: for all (i,t),

$$\Pr\left(Y_{it} = y | (X_{is})_{s \leq t}, \mathcal{F}\right) = g(y; X_{it}'\beta_i + \alpha_i + \xi_t), \tag{2}$$

with, almost surely, $\sup_i \|\beta_i\| \leq C_\beta < \infty$. Moreover, g is unknown.

(b) Monotonicity and smoothness: there exists $\bar{y} \in \mathcal{Y}$ such that the function $v \mapsto g(\bar{y}; v)$ is strictly increasing and L-Lipschitz.

(c) Cross-section independence and weak serial dependence:

1. *Conditional on $\mathcal{F}$, $\{(Y_{it}, X_{it}') : t = 1, 2, \ldots\}$ is independent across i.*
2. *Let $\mu > 1$. Conditional on $\mathcal{F}$, for each i, $\{(Y_{it}, X_{it}') : t = 1, 2, \ldots\}$ is α-mixing with mixing coefficient satisfying $\sup_i a_i(m) = O(m^{-\mu})$ as $m \to \infty$, where*

$$a_i(m) \equiv \sup_t \sup_{A \in \mathcal{A}_t^i, B \in \mathcal{B}_{t+m}^i} \left| \Pr(A \cap B) - \Pr(A) \Pr(B) \right|,$$

$\mathcal{A}_t^i$ is the σ-field generated by $((Y_{it}, X_{it}'), (Y_{it-1}, X_{it-1}'), \ldots)$, and $\mathcal{B}_t^1$ is the σ-field generated by $((Y_{it}, X_{it}'), (Y_{it+1}, X_{it+1}'), \ldots)$.

(d) Conditional on $\mathcal{F}$, X_{it} has density p_{it} with respect to the Lebesgue measure on $\mathbf{R}^K$ that satisfies $p_{it}(x) \leq p_{max} < \infty$ for all (i, t) and $x \in \mathbf{R}^K$.

(e) Let K denote a bounded kernel function. For all strictly monotonic functions $f : \mathbf{R} \to (0, 1)$ and $x = (x^{(1)}, x^{(2)}) \in \mathcal{X}$, almost surely, there exists a constant $c_{f,x^{(2)}}$ nontrivially depending on f such that, for all i,

$$\frac{\frac{1}{h_T T} \sum_{t=1}^T K\left(\frac{X_{it} - x}{h_T}\right) f(\xi_t)}{\frac{1}{h_T T} \sum_{t=1}^T K\left(\frac{X_{it} - x}{h_T}\right)} \to c_{f,x^{(2)}} \text{ as } T \to \infty.$$

Assumptions 1(a) and 1(b) define the class of two-way fixed effects models we focus on. In Assumption 1(a), the distribution of outcome Y_{it} depends on the observed characteristics X_{it} and fixed effects $(\alpha_i, \beta_i, \xi_t)$ via a single index and an unknown link function g. Individual-specific slopes β_i capture heterogeneous effect of X_{it} and are potentially unobserved to the researcher.[3] In Assumption 1(b), we impose monotonicity and smoothness conditions on the dependence of g on the single index v at some known $\bar{y} \in \mathcal{Y}$.[4] Most link functions, e.g., logit, probit, and Poisson, satisfy these conditions.

Assumptions 1(c)1 and 1(c)2 impose dependence restrictions across individual and time dimensions of the panel; both are standard in the panel data literature

[3]One can use an *it*-specific β_{it} and specify $\beta_{it} = \gamma r_{it}$ to capture observed heterogeneity in slopes, where r_{it} is a vector of observed characteristics of individual i at time t. This is equivalent to adding $x_{it} r_{it}$ in (2) with common slopes γ across individuals and time periods.

[4]If $v \mapsto g(\bar{y}; v)$ is strictly decreasing, one can transform $((\alpha_i, \beta_i, \xi_t)_{i,t}, g(\cdot;\cdot))$ to $((-\alpha_i, -\beta_i, -\xi_t)_{i,t}, g(\cdot;-\cdot))$ that delivers the same model (2) and $v \mapsto g(\bar{y}; -v)$ is strictly increasing in v.

(e.g., Assumption 4.1 in Fernández-Val and Weidner (2016)). Assumption 1(c)1 requires cross-sectional independence across individuals conditional on common shocks. Assumption 1(c)2 requires $\alpha-$mixing properties across time periods conditional on common shocks. It does not impose identical nor stationary distribution over the time dimension. The model allows for lagged outcomes as explanatory covariates, though Assumption 1(c)2 may rule out some forms of dynamics.[5]

Assumption 1(e) requires the observed individual time series to have a Birkhoff's almost-sure ergodic property. It is sufficient for constructing the limiting objects from the individual time series so that their true values are known to the econometrician in the setting of identification discussion (i.e., when T and N are infinity). When, for all i, the corresponding time series $\{(\xi_t, X_{it})\}_{t=2,3,...}$ are strictly stationary with strong mixing conditions (see Hansen (2008)), this assumption holds as long as the distribution of $\xi_t|X_{it} = x$ does not depend on $x_{it}^{(1)}$, an exogeneity condition that can be qualified in an applied setting. It allows for ξ_t and $\xi_{t'}$ to be correlated as long as the correlation vanishes as the time periods are distant enough. It also accommodates some non-stationary $(\xi_t)_t$ such as deterministic time trends. For instance, if $\xi_t = t - 1$, then $\{f(\xi_t)\}_{t=1,2,...}$ is a bounded strictly monotonic sequence of real numbers with limit in $\{0, 1\}$ so that $c_{f,x^{(2)}} \in \{0, 1\}$. Another example is periodic time trend: there exists T_0 such that $\xi_{t+T_0} = \xi_t$ non-random. Then, $c_{f,x^{(2)}} \equiv c_f$.

Lastly, Assumption 1(d) focuses on continuous covariates and rules out discrete ones. This choice is mainly to simplify the exposition. In the presence of covariates whose distributions are mixtures of discrete and continuous distributions with known dominating measure, one can accordingly modify Assumption 1(e) such that the conditional expectation at point masses is identified in the limit by an empirical frequency justified by the law of large numbers, possibly combined with kernel smoothing.

In Appendix E, we present two extensions of model (2). The first one is a model with time-specific slopes:

$$\Pr\left(Y_{it} = y|(X_{is})_{s\leq t}, \mathcal{F}\right) = g(y; X_{it}'\beta_t + \alpha_i + \xi_t), \tag{3}$$

where β_t captures potentially heterogeneous effect of X_{it} across time periods. The

[5]For instance, Andrews (1984) discussed simple autoregressive models that are not strongly mixing. The nonlinearity in (2) makes it more difficult to link the regressive coefficient to the α-mixing coefficient and verify the mixing property.

second one is with multinomial outcomes:

$$y_{it} = \underset{j=1,\ldots,J}{\arg\max} \quad \left\{\alpha_{ij} + \xi_{tj} + x'_{tj}\beta_{ij} - u_{itj}\right\}, \tag{4}$$

where $(u_{it1}, \ldots, u_{itJ})$ are independent of $(\alpha_{ij}, \xi_{tj}, \beta_{ij}, x_{tj})_{j=1}^{J}$ and distributed according to density g^*. Define $v_{itj} = \alpha_{ij} + \xi_{tj} + x'_{tj}\beta_{ij}$. Then,

$$g(y; v_{it1}, \ldots, v_{itJ}) = \sum_{j=1}^{J} \mathbf{1}\{y = j\} \Pr(u_{itj} - u_{itj'} \leq v_{itj} - v_{itj'}, \text{ for any } j' \neq j),$$

where the right-hand side is a function of J indexes $v_{it} = (v_{itj})_{j=1}^{J}$ and J is known.

Given the data generating process described by Assumption 1, the econometrician observes a finite sample of size NT from the infinite population. Denote by M_{it}^{NT} the indicator of (i,t) belonging to this finite sample, i.e., (i,t) is in this finite sample if and only if $M_{it}^{NT} = 1$. Moreover, $\sum_{i=1}^{\infty}\sum_{t=1}^{\infty} M_{it}^{NT} = NT$ and $M_{it}^{NT} \in \{0,1\}$ for all i,t. The sequence of infinite binary arrays $M^{NT} \equiv (M_{it}^{NT})_{i,t}$ with NT positive entries governs the sampling process.

Assumption 2 (Sampling)**.**

(a) Independent sampling: $M^{NT} \perp ((Y_{it}, X'_{it})_{i,t}, \mathcal{F})$.

(b) Single increasing panel: $[N \leq \tilde{N} \text{ and } T \leq \tilde{T}] \implies [M_{it}^{NT} \leq M_{it}^{\tilde{N}\tilde{T}}, \quad \forall (i,t)]$.

(c) Balanced NT-panels: for all N, T, i, t,

$$\sum_{i=1}^{\infty}\sum_{t=1}^{\infty} M_{it}^{NT} = NT$$

and

$$M_{it}^{NT} = 1 \implies \begin{cases} M_{is}^{NT} = 1, & \forall s : \exists j, M_{js}^{NT} = 1, \\ M_{jt}^{NT} = 1, & \forall j : \exists s, M_{js}^{NT} = 1. \end{cases}$$

Assumption 2(a) rules out any dependence between the sampling process and the joint distribution of population outcomes and latent shocks. This is a "Missing-At-Random" assumption on the infinite population. Assumption 2(b) is analogous to the assumption of "staggered adoption" in the causal difference-in-differences (DID) literature: once a cell (i,t) has entered the sample ("treatment" in the DID literature), it remains in subsequent ones. By Assumption 2(c), we consider

balanced panels in the subsequent analysis to simplify the exposition. One can adapt our results and the proofs to allow for unbalanced sampling processes.

Assumptions 1 and 2 provide an asymptotic framework within which the identification of individual or time-specific parameters $(\alpha_i, \beta_i, \xi_t)$ is meaningful. Usually speaking, these parameters depend on sample size and their triangular sequence may not have a meaningful limit.[6] Our framework solves this problem by assuming a well-defined infinite population (Assumption 1) and sequences of increasing consistent samples (Assumption 2). Whether a cross-sectional unit i or time period t appears in observed samples only depends on the sampling process M^{NT} which we assume independent from outcomes and latent shocks.

3 Identification and Estimation

To simplify the exposition, we consider the following rectangular sampling process that satisfies Assumption 2:

$$M_{it}^{NT} = \begin{cases} 1 & \text{if } i \leq N, t \leq T \\ 0 & \text{otherwise.} \end{cases}$$

and let $N, T \to \infty$. We are interested in identifying and estimating $(\alpha_i, \beta_i)_{i=1,2,...}$, $(\xi_t)_{t=1,2,...}$, and function g in model (2).

3.1 Identification

We present the arguments for $X_{it} = (X_{it}^{(1)}, X_{it}^{(2)}) \in \mathcal{X} \subset \mathbf{R}^2$. We extend the identification results to the cases of time-specific slopes (3) and multinomial outcomes (4) in Appendix E.

For any i such that $\beta_i^{(1)} \neq 0$, we first define:

$$z_{i \to i'}(x^{(1)}; x^{(2)}) \equiv [x^{(1)}\beta_{i'}^{(1)} + x^{(2)}(\beta_{i'}^{(2)} - \beta_i^{(2)}) + \alpha_{i'} - \alpha_i]/\beta_i^{(1)}. \tag{5}$$

Intuitively, $z_{i \to i'}(x^{(1)}; x^{(2)})$ is interpreted as a *compensating variable*, i.e., the needed value of $x^{(1)}$ for individual i with $x^{(2)}$ to make her and i''s indexes equal:

[6] Instead, if some kind of uniform convergence holds, e.g., $\lim_{N,T\to\infty} \sup_{i\leq N, t\leq T} \|(\hat{\alpha}_i, \hat{\beta}_i, \hat{\xi}_t) - (\alpha_i^0, \beta_i^0, \xi_t^0)\| \xrightarrow{p} 0$ where $(\hat{\alpha}_i, \hat{\beta}_i, \hat{\xi}_t)$ are estimators of the true ones $(\alpha_i^0, \beta_i^0, \xi_t^0)$ when the sample size is NT, then one could identify the distributional features of the fixed effects.

$\alpha_i + \xi_t + \beta_i^{(1)} z_{i\to i'}(x^{(1)}; x^{(2)}) + \beta_i^{(2)} x^{(2)} = \alpha_{i'} + \xi_t + \beta_{i'}^{(1)} x^{(1)} + \beta_{i'}^{(2)} x^{(2)}$. The following definitions formalize the idea of compensation in the infinite population.

Definition 1 (Compensable)**.** *Individual i' is said to be compensable by individual i at point $(x^{(1)}, x^{(2)})$ if and only if $(z_{i\to i'}(x^{(1)}; x^{(2)}), x^{(2)}) \in \mathcal{X}_i$.*

Note that if $\beta_i^{(1)} = 0$, then by definition individual i cannot compensate any other individuals. In Appendix A, we show that the set of individuals with $\beta_i^{(1)} = 0$, denoted by $\mathcal{I}_0$, is identified under Assumptions 1 and 2. For those in $\mathcal{I}_0$ that are compensable, we can identify their parameters using the same arguments in our main results. Instead, for those in $\mathcal{I}_0$ that are not compensable, our identification arguments do not apply. In the remaining part of the paper, we will focus on the subpopulation with non-zero $\beta_i^{(1)}$, i.e., $\mathbb{N} \setminus \mathcal{I}_0$.

Definition 2 (Compensating network)**.** *Let $\mathcal{G}^\infty$ denote the* compensating network *with an edge between i to j with $i, j \in \mathbb{N} \setminus \mathcal{I}_0$, denoted $i \longleftrightarrow j$, if and only if either individual j is compensable by individual i at least at $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_j$ for $k = 1, 2, 3$, with*

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix}$$

being nonsingular, or i is compensable by individual j at least at $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_i$ for $k = 1, 2, 3$, with the same rank condition.

In general, $\mathcal{G}^\infty$ induces a partition of $\mathbb{N} \setminus \mathcal{I}_0$ that contains at most a countably many disjoint subsets of $\mathbb{N} \setminus \mathcal{I}_0$. Within each subset, two individuals i_1 and i_l are connected via a sequence $(i_1, i_2), \ldots, (i_{l-1}, i_l)$ in which either i_{k-1} is compensable by i_k or the other way around for any $k = 2, \ldots, l$. It is possible that some subsets are singletons, i.e., each of them contains only one individual that neither compensate nor can be compensated by others. The next theorem states a relative identification result within each subset that contains at least two individuals. The proof can be found in Appendix A.

Theorem 1 (Relative identification)**.** *Suppose that Assumptions 1 and 2 hold. Denote by $\{\mathcal{I}_r : r = 1, 2, \ldots\}$ the partition of $\mathbb{N} \setminus \mathcal{I}_0$ induced by $\mathcal{G}^\infty$. Then, $(\alpha_i - \alpha_j)/\beta_j^{(1)}$, $\beta_i^{(1)}/\beta_j^{(1)}$, and $(\beta_i^{(2)} - \beta_j^{(2)})/\beta_j^{(1)}$ are identified for any $i, j \in \mathcal{I}_r$ and any r with $\mathcal{I}_r$ containing at least two individuals.*

Intuitively, for a sequence $(i_1, i_2), \ldots, (i_{l-1}, i_l)$ with $i_k \in \mathcal{I}_r$ for $k = 1, \ldots, l$, if individual i_{k-1} is compensable by i_k at the three points in Definition 2, one can then identify $z_{i_k \to i_{k-1}}(x^{(1)1}; x^{(2)1})$, $z_{i_k \to i_{k-1}}(x^{(1)2}; x^{(2)2})$, and $z_{i_k \to i_{k-1}}(x^{(1)3}; x^{(2)3})$ by comparing i_{k-1}'s and i_k's choices, $y_{i_{k-1}t}$ and $y_{i_k t}$, over time. The rank condition in Definition 2 ensures the unique recovery of $(\alpha_{i_{k-1}} - \alpha_{i_k})/\beta_{i_k}^{(1)}$, $\beta_{i_{k-1}}^{(1)}/\beta_{i_k}^{(1)}$, and $(\beta_{i_{k-1}}^{(2)} - \beta_{i_k}^{(2)})/\beta_{i_k}^{(1)}$ from $z_{i_k \to i_{k-1}}(x^{(1)1}; x^{(2)1})$, $z_{i_k \to i_{k-1}}(x^{(1)2}; x^{(2)2})$, and $z_{i_k \to i_{k-1}}(x^{(1)3}; x^{(2)3})$. In essence, the rank requirement rules out the situation in which one point, say $(x^{(1)3}, x^{(2)3})$, lies on the line defined by $(x^{(1)1}, x^{(2)1})$ and $(x^{(1)2}, x^{(2)2})$. When $(x^{(1)}, x^{(2)}) \in \mathcal{X}_{i_{k-1}}$ are continuous and the set of points at which i_{k-1} is compensable by i_k has positive Lebesgue measure, the rank condition automatically holds. One can apply the same reasoning to the case in which i_k is compensable by i_{k-1} at the three points in Definition 2 and to all the pairs in the sequence. Consequently, we achieve the relative identification in Theorem 1.

Theorem 1 relates the ability of identifying individual-specific parameters in model (2) to the connectedness of $\mathcal{G}^\infty$, an insight that joins some recent literature using "overlapping graphs" as a key identifying device (see, e.g., Abowd et al., 1999; Jochmans and Weidner, 2019; Lei and Ross, 2024). To see this point, suppose that $\mathcal{G}^\infty$ contains two connected components that contains individual 1 and 2, respectively. According to Theorem 1, we can identify $((\alpha_i - \alpha_1)/\beta_1^{(1)}, \beta_i^{(1)}/\beta_1^{(1)}, (\beta_i^{(2)} - \beta_1^{(2)})/\beta_1^{(1)})$ for $i \in \mathcal{I}_1$ and $((\alpha_j - \alpha_2)/\beta_2^{(1)}, \beta_j^{(1)}/\beta_2^{(1)}, (\beta_j^{(2)} - \beta_2^{(2)})/\beta_2^{(1)})$ for any $j \in \mathcal{I}_2$. However, the magnitude of $(\alpha_1, \beta_1^{(1)})$ relative to $(\alpha_2, \beta_2^{(1)})$ is not identified. As a result, the magnitudes of $(\alpha_i, \beta_i^{(1)}, \beta_i^{(2)})$ relative to $(\alpha_j, \beta_j^{(1)}, \beta_j^{(2)})$ for $i \in \mathcal{I}_1$ and $j \in \mathcal{I}_2$ are not identified. In contrast, if $\mathcal{G}^\infty$ has only one component and is connected, we then identify $(\alpha_i, \beta_i^{(1)}, \beta_i^{(2)} - \beta_1^{(2)})$ relative to $(\alpha_1, \beta_1^{(1)})$ for all $i \in \mathbb{N} \setminus \mathcal{I}_0$.

What determines the connectedness of $\mathcal{G}^\infty$ is the support of $X_{it}^{(1)}$, $\mathcal{X}_i^1$, for $i \in \mathbb{N} \setminus \mathcal{I}_0$. When $\mathcal{X}_i$ has a large support, i.e., $\mathcal{X}_i^1 = \mathbb{R}$, any other individual is then compensable by i at any point. $\mathcal{G}^\infty$ is therefore a connected network. Nevertheless, depending on the supports of other individuals, the large support condition may be unnecessary for having a connected $\mathcal{G}^\infty$. For instance, suppose that $\mathcal{X}_i$ is uniformly bounded for $i \geq 2$. Because of Assumption 1(a), the required compensation for $i \geq 2$ is also uniformly bounded. As long as $\mathcal{X}_1^1$ is larger than this uniformly bounded set of compensation, $\mathcal{G}^\infty$ will be connected. In Appendix C, we provide two examples along the lines of this reasoning and illustrate how economic restrictions help alleviate the support requirement.

In the next assumption, we suppose a star structure in $\mathcal{G}^\infty$ and propose conditions to identify remaining parameters. Let $\mathrm{P}_v(z, x^{(2)}) = (z, x^{(2)})v$ be the operation of inner product.

Assumption 3 (Identification)**.**

(a) There exists $i^\star \in \mathbb{N} \setminus \mathcal{I}_0$ such that for all $j \in \mathbb{N} \setminus \mathcal{I}_0$, there exists an edge $i^\star \longleftrightarrow j$ in $\mathcal{G}^\infty$.

(b) Let $\mathcal{Z}_{it}^{i^\star}$ denotes the support of $(z_{i^\star \to i}(x_{it}^{(1)}, x_{it}^{(2)}), x_{it}^{(2)})$ and $\mathcal{Z}_t^{i^\star} = \cap_{i \in \mathbb{N} \setminus \mathcal{I}_0} \mathcal{Z}_{it}^{i^\star}$. For some $(t, r) \in \mathbb{N} \times \mathbf{R}$, $\left\{ z \in \mathcal{Z}_t^{i^\star} : \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})}(z) = r \right\}$ is not a singleton.

The star structure in Assumption 3(a) implies that $\mathcal{G}^\infty$ is connected. As discussed previously, we can then identify $(\alpha_i, \beta_i^{(1)}, \beta_i^{(2)} - \beta_{i^\star}^{(2)})$ relative to $(\alpha_{i^\star}, \beta_{i^\star}^{(1)})$ for any $i \in \mathbb{N} \setminus \mathcal{I}_0$. Assumption 3(b) gives the condition under which $z_{i^\star}(x^{(1)}; x^{(2)})$ compensates between $x^{(1)}$ and $x^{(2)}$ for $i^\star$. It is used to identify $\beta_{i^\star}^{(2)}$ and $\beta_i^{(2)}$ relative to $\beta_{i^\star}^{(1)}$ for $i \in \mathbb{N} \setminus \mathcal{I}_0$.

Theorem 2. *Suppose that Assumptions 1–3 hold. Then,*

- $\left(\frac{\alpha_i - \alpha_{i^\star}}{\beta_{i^\star}^{(1)}}, \frac{\beta_i^{(1)}}{\beta_{i^\star}^{(1)}}, \frac{\beta_i^{(2)}}{\beta_{i^\star}^{(1)}} \right)$ *are identified for any $i \in \mathbb{N} \setminus \mathcal{I}_0$.*
- $\frac{\xi_t - \xi_s}{\beta_{i^\star}^{(1)}}$ *is identified for all $s, t \in \mathbb{N}$ such that*

$$\left(\cap_{i \in \mathbb{N} \setminus \mathcal{I}_0} \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})}(\mathcal{Z}_{is}^{i^\star}) + \xi_s \right) \cap \left(\cap_{i \in \mathbb{N} \setminus \mathcal{I}_0} \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})}(\mathcal{Z}_{it}^{i^\star}) + \xi_t \right) \neq \emptyset. \quad (6)$$

- *For $t^\star$, let*

$$\mathcal{T}^\star = \{ t : \exists\ (t^\star, t_1), (t_1, t_2), \ldots, (t_l, t) \text{ such that } (6) \text{ holds for each pair in the sequence} \}.$$

Then, $g(y; \beta_{i^\star}^{(1)} u + \alpha_{i^\star} + \xi_{t^\star})$ is identified for any $(y, u) \in \mathcal{Y} \times \cup_{t \in \mathcal{T}^\star} \left(\cap_{i \in \mathbb{N} \setminus \mathcal{I}_0} \mathrm{P}_{(1, \beta_{i^\star}^{(2)} / \beta_{i^\star}^{(1)})}(\mathcal{Z}_{it}^{i^\star}) + (\xi_t - \xi_{t^\star}) / \beta_{i^\star}^{(1)} \right)$.

In a similar vein to Theorem 1, the second result of Theorem 2 identifies relatively ξ_t by compensating the difference $\xi_t - \xi_s$ with $(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})(z_s - z_t)$ for some $z_s \in \cap_{i \in \mathbb{N} \setminus \mathcal{I}_0} \mathcal{Z}_{is}^{i^\star}$ and $z_t \in \cap_{i \in \mathbb{N} \setminus \mathcal{I}_0} \mathcal{Z}_{it}^{i^\star}$. The overlapping-support condition in (6) is sufficient for applying this compensation argument. In the third result of Theorem 2, the set $\mathcal{T}^\star$ gathers time periods that are connected to $t^\star$ and the corresponding time-fixed effects are identified relative to $\xi_{t^\star}$. Together with the

identification of $\left(\frac{\alpha_i-\alpha_{i^\star}}{\beta_{i^\star}^{(1)}}, \frac{\beta_i^{(1)}}{\beta_{i^\star}^{(1)}}, \frac{\beta_i^{(2)}}{\beta_{i^\star}^{(1)}}\right)$, this result implies the identification of single index $u \in \cup_{t\in\mathcal{T}^\star}\left(\cap_{i\in\mathbb{N}\setminus\mathcal{I}_0} \mathrm{P}_{(1,\beta_{i^\star}^{(2)}/\beta_{i^\star}^{(1)})}(\mathcal{Z}_{it}^{i^\star}) + (\xi_t - \xi_{t^\star})/\beta_{i^\star}^{(1)}\right)$. We then identify $g(y; \beta_{i^\star}^{(1)}u + \alpha_{i^\star} + \xi_{t^\star})$ as a function of any $y \in \mathcal{Y}$ and u in this support.

Normalization. Suppose that $\mathcal{T}^\star = \mathbb{N}$ in Theorem 2. We then identify $\frac{\xi_t-\xi_{t^\star}}{\beta_{i^\star}^{(1)}}$ for all $t \in \mathbb{N}$. Denote $\tilde{\beta}_i = \beta_i/\beta_{i^\star}^{(1)}$, $\tilde{\alpha}_i = (\alpha_i - \alpha_{i^\star})/\beta_{i^\star}^{(1)}$, $\tilde{\xi}_t = (\xi_t - \xi_{t^\star})/\beta_{i^\star}^{(1)}$, and $\tilde{g}(y; u_{it}) = g(y; \beta_{i^\star}^{(1)}u_{it} + \alpha_{i^\star} + \xi_{t^\star})$ where $u_{it} = x_{it}'\beta_i/\beta_{i^\star}^{(1)} + (\alpha_i - \alpha_{i^\star})/\beta_{i^\star}^{(1)} + (\xi_t - \xi_{t^\star})/\beta_{i^\star}^{(1)}$. Note that $((\alpha_i, \beta_i, \xi_t)_{i,t}, g)$ and $((\tilde{\alpha}_i, \tilde{\beta}_i, \tilde{\xi}_t)_{i,t}, \tilde{g})$ deliver the same model in (2). When $i = i^\star$, we have $u_{i^\star t} = x_{i^\star t}$ and identify $\tilde{g}(\bar{y}; x_{i^\star t}) = g(\bar{y}; \beta_{i^\star}^{(1)}x_{i^\star t} + \alpha_{i^\star} + \xi_{t^\star})$ as a function of $x_{i^\star t}$ where $\bar{y}$ in defined in Assumption 1(b). Then, because $g(\bar{y}; v)$ is strictly increasing in v (Assumption 1(b)), we can identify the sign of $\beta_{i^\star}^{(1)}$ (and of $\beta_i^{(k)}$ for any i and k). Consequently, one can normalize $\alpha_{i^\star} = 0$, $\beta_{i^\star}^{(1)} = \mathbf{1}\{\beta_{i^\star}^{(1)} > 0\} - \mathbf{1}\{\beta_{i^\star}^{(1)} < 0\}$ (or equivalently $\beta_i^{(k)} = \mathbf{1}\{\beta_i^{(k)} > 0\} - \mathbf{1}\{\beta_i^{(k)} < 0\}$ for some i and k), and $\xi_{t^\star} = 0$ without loss of generality.

Despite the normalization, Theorem 2 implies the identification of marginal effects of $x_{it} = x$ for $y_{it} = y$, $\frac{\partial g(y;x'\beta_i+\alpha_i+\xi_t)}{\partial x}$, by $\frac{\partial \tilde{g}(y;x'\beta_i+\alpha_i+\xi_t)}{\partial x} = \frac{\partial \tilde{g}(y;x'\beta_i+\alpha_i+\xi_t)}{\partial u}\tilde{\beta}_i$, and its average over time for individual i by $\tilde{\beta}_i \mathbb{E}_{\tilde{\xi},x|\tilde{\alpha}_i,\tilde{\beta}_i}\left[\frac{\partial \tilde{g}(y;x'\tilde{\beta}_i+\tilde{\alpha}_i+\tilde{\xi})}{\partial u}\right]$. Both are typically the objects of interest in applied research. Different from those in models with homogeneous slopes, the marginal effects in model (2) can vary across individuals due to heterogeneous slopes $\tilde{\beta}_i$. Our identification results enable to quantify the extent to which the slope heterogeneity explains the dispersion in marginal effects, a point we will investigate in the empirical illustration in Section 5.

3.2 Consistency of Sieve MLE

The identification of $(\alpha_{i0}, \beta_{i0}, \xi_{t0})_{i\in\mathbf{N},t\in\mathbf{T}}$ and the link function g_0 (where subscript 0 denotes the true parameter values) in Theorem 2 hints on the potential of consistently estimating these parameters in a large-N-and-large-T asymptotic framework. In this section, we propose a consistency result for the sieve maximum likelihood estimation (MLE) of $(\alpha_{i0}, \beta_{i0}, \xi_{t0})_{i\in\mathbf{N},t\in\mathbf{T}}$ and g_0, a natural nonparametric extension of the fixed effects MLE routinely used in the literature. Its implementation is similar to the parametric MLE and we discuss some novelties in Section 3.3, e.g., shape restrictions on the link function.

We assume that $\mathcal{G}^\infty$ is connected and $\mathcal{T}^\star = \mathbb{N}$ so that we can normalize $(\alpha_{i^\star 0}, \beta_{i^\star 0}^{(1)}) = (0, 1)$ (or $(\alpha_{i^\star 0}, \beta_{i^\star 0}^{(1)}) = (0, -1)$) and $\xi_{t^\star 0} = 0$ for some $i^\star$ and $t^\star$.

Accordingly, the parameter space defined in the estimation takes into account these normalizations. Suppose $\mathcal{Y}$ is a set of finite outcomes $\{0, 1, \ldots, L\}$.[7] Define $h_y(v) := \log\left(g(y; v)/g(0; v)\right)$ for any $y \in \mathcal{Y}$. In the case of logistic link function, $h_1(v) = v$. It is theoretically equivalent to use $(h_y(\cdot))_{y=1}^L$ and $(g_y(\cdot))_{y=0}^L$ under the restrictions $g_y(\cdot) > 0$ for any $y = 0, \ldots, L$ and $\sum_{y=0}^L g_y(\cdot) = 1$. It is more of a practical matter to consider sieve estimates of $(h_y(\cdot))_{y=1}^L$ because they are free of the aforementioned restrictions.

Denote by $\Omega_{\alpha,\beta}$ and Ω_ξ the support of $(\alpha_{i0}, \beta_{i0})$ and ξ_{t0}, respectively. We propose the following regularity conditions on the distribution of $(\alpha_{i0}, \beta_{i0})$ and ξ_{t0}.

Assumption 4 (Fixed effects)**.** *$(\alpha_{i0}, \beta_{i0})$ and ξ_{t0} are continuous random variables in compact domains $\Omega_{\alpha,\beta} \subset \mathbb{R}^3$ and $\Omega_\xi \subset \mathbb{R}$, respectively. The density functions of $(\alpha_{i0}, \beta_{i0})$ and ξ_{t0} are uniformly bounded away from zero by some $c > 0$. $(\alpha_{i0}, \beta_{i0})$ are independent across i. Let $\mu > 1$. $\{\xi_t : t = 1, 2, \ldots\}$ is α-mixing with mixing coefficient satisfying $a(m) = O(m^{-\mu})$ as $m \to \infty$, where*

$$a(m) \equiv \sup_t \sup_{A \in \mathcal{A}_t^i, B \in \mathcal{B}_{t+m}^i} \left|\Pr(A \cap B) - \Pr(A)\Pr(B)\right|,$$

and $\mathcal{A}_t^i$ is the σ-field generated by $(\xi_t, \xi_{t-1}, \ldots)$, and $\mathcal{B}_t^1$ is the σ-field generated by $(\xi_t, \xi_{t+1}, \ldots)$. Moreover, $(X_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0})$ is identically distributed across i and t.

Assumption 4 requires bounded support on $(\alpha_{i0}, \beta_{i0})$ and ξ_{t0}, similarly to the support condition on β_{i0} in Assumption 1(a). It also implies that the density $p_{it}(x)$ in Assumption 1(d) is written as $p_{it}(x) = f_{0x}(x|\alpha_{i0}, \beta_{i0}, \xi_{t0})$ where $f_{0x}(x|\alpha, \beta, \xi)$ is the density of X_{it} conditional on $(\alpha_{i0}, \beta_{i0}, \xi_{t0}) = (\alpha, \beta, \xi)$. The $\alpha-$mixing property on $(\xi_t)_t$ imposes a weak serial dependence on $(\xi_t)_{t \geq 1}$. It is compatible with the ergodic requirement in Assumption 1(e).
First, denote the log-likelihood function by

$$\begin{aligned} \mathscr{L}_{NT}(\theta^{NT}) &:= \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \log g(y_{it}; \alpha_i + \xi_t + x_{it}'\beta_i) \\ &= \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \log\left[\frac{\exp\{h_{y_{it}}(\alpha_i + \xi_t + x_{it}'\beta_i)\}}{1 + \sum_{y=1}^L \exp\{h_y(\alpha_i + \xi_t + x_{it}'\beta_i)\}}\right] \end{aligned} \tag{7}$$

where $\theta^{NT} = ((\alpha_i, \beta_i)_{i=1}^N, (\xi_t)_{t=1}^T, h)$ and $h = (h_y)_{y \in \mathcal{Y}, y \neq 0}$.

[7] For the case of infinitely countably many outcomes, one can transform the model to a truncated one with finitely many outcomes.

Define the sieve maximum likelihood estimators as:

$$\hat{\theta}^{NT} := (\hat{\theta}_1^{NT}, \underbrace{\hat{h}_1^{NT}, \ldots, \hat{h}_L^{NT}}_{=:\hat{h}^{NT}}) = \underset{\theta^{NT} \in \Theta_1^{NT} \times \Theta_h^{NT}}{\arg\max} \mathscr{L}_{NT}(\theta^{NT}), \tag{8}$$

where the compact set Θ_1^{NT} contains the true $\theta_{10}^{NT} = (\alpha_{i0}, \beta_{i0}, \xi_{t0})_{i=1,\ldots,N;t=1,\ldots,T}$. The set Θ_h^{NT} is a subset of $\Theta_h = \Theta_{h_1} \times \cdots \times \Theta_{h_L}$ with Θ_{h_l} containing the true value h_{l0} for $l = 1, \ldots, L$. Denote by $\theta_0^{NT} = (\theta_{10}^{NT}, h_0)$. We need the following assumption on the sieve space for the consistency of $\hat{h}^{NT}$.

Assumption 5 (Sieve space)**.** *The closure of Θ_h, $\bar{\Theta}_h$, is compact in the relative topology generated by some norm $\|\cdot\|_h$. Moreover, $\cup_{T,N}\Theta_h^{NT}$ is a dense subset of $\bar{\Theta}_h$ with respect to $\|\cdot\|_h$, and $\Theta_h^{NT} \subset \Theta_h^{\tilde{N}\tilde{T}}$ if $N \leq \tilde{N}$ and $T \leq \tilde{T}$.*

Assumption 5 is standard in the sieve literature. Examples of Θ_h and Θ_h^{NT} include Hölder class of functions and linear sieve spaces such as polynomials and splines. We refer to Gallant and Nychka (1987), Chen (2007) and Freyberger and Masten (2019) among others for more examples of $\bar{\Theta}_h$ and Θ_h^{NT}. In Section 3.3, we discuss some practical issues related to the choice of sieve space in the context of (8).

The challenge in establishing the consistency of $\hat{\theta}_1^{NT}$ is its increasing dimensionality. The conventional Euclidean distance measure may not be suitable because of a lack of invariance with respect to the increasing dimension of $\hat{\theta}_1^{NT}$. The max norm is invariant to the increasing dimensionality, but the limiting spaces ($\Omega_{\alpha,\beta}^{\mathbb{N}}$ and $\Omega_{\xi}^{\mathbb{N}}$) are noncompact under this norm.

To circumvent this challenge, we reformulate $\hat{\theta}_1^{NT}$ as a collection of mappings rather than point estimates. In fact, under appropriate regularity conditions (Assumption 7 below), $\hat{\theta}^{NT}$ is asymptotically close to the maximizer of

$$\mathscr{L}_{NT}^0(\theta_1^{NT}, h) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{E}_0\left[\log\left(\frac{\exp\{h_{y_{it}}(\alpha_i + \xi_t + x_{it}'\beta_i)\}}{1 + \sum_{y=1}^{L}\exp\{h_y(\alpha_i + \xi_t + x_{it}'\beta_i)\}}\right)\middle| x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0}\right],$$

where $\mathbb{E}_0[\cdot|x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0}]$ refers to the expectation with respect to y_{it} conditional on $(x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0})$ and given h_0. We then can rewrite $\mathscr{L}_{NT}^0(\hat{\theta}_1^{NT}, \hat{h})$ as:

$$\mathscr{L}^0(\hat{G}^{NT}, \hat{h}^{NT}; F_0^{NT}) := \int \mathbb{E}_0\left[\log\left(\frac{\exp\{\hat{h}_Y^{NT}(a + e + x'b)\}}{1 + \sum_{y=1}^{L}\exp\{\hat{h}_y^{NT}(a + e + x'b)\}}\right)\middle| x, \alpha, \beta, \xi\right] d\hat{G}_1^{NT}(a, b|\alpha, \beta) d\hat{G}_2^{NT}(e|\xi) dF_0^{NT}(x, \alpha, \beta, \xi), \tag{9}$$

where F_0^{NT} is the empirical distribution of $(x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0})_{1\leq i\leq N; 1\leq t\leq T}$:

$$F_0^{NT}(x,\alpha,\beta,\xi) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\text{Dirac}_{x_{it}}(x)\text{Dirac}_{(\alpha_{i0},\beta_{i0})}(\alpha,\beta)\text{Dirac}_{\xi_{t0}}(\xi),$$

and $\hat{G}^{NT} = (\hat{G}_1^{NT}, \hat{G}_2^{NT})$. Besides, $\hat{G}_1^{NT}$ is a mapping from $(\alpha,\beta) \in \Omega_{\alpha,\beta}$ to a probability distribution in $\Omega_{\alpha,\beta}$, $\hat{G}_1^{NT}(\cdot|\alpha,\beta)$ with $\hat{G}_1^{NT}(\cdot|\alpha_{i0},\beta_{i0}) = \text{Dirac}_{(\hat{\alpha}_i^{NT},\hat{\beta}_i^{NT})}$ for $1 \leq i \leq N$, and $\hat{G}_2^{NT}$ is a mapping from $\xi \in \Omega_\xi$ to a probability distribution in Ω_ξ, $\hat{G}_2^{NT}(\cdot|\xi)$ with $\hat{G}_2^{NT}(\cdot|\xi_{t0}) = \text{Dirac}_{\hat{\xi}_t^{NT}}$ for $1 \leq t \leq T$. Intuitively, if $(\hat{\alpha}_i^{NT}, \hat{\beta}_i^{NT})$ converges to (α_{i0},β_{i0}) in probability for any i, the mapping $\hat{G}_1^{NT}$ will then "converge" to the embedding from $(\alpha,\beta) \in \Omega_{\alpha,\beta}$ to $\text{Dirac}_{(\alpha,\beta)}$, denoted by id_1. Conversely, if $\hat{G}_1^{NT}$ converges to id_1, we then obtain the consistency of $(\hat{\alpha}_i^{NT}, \hat{\beta}_i^{NT})$. Similarly, to obtain the consistency of $\hat{\xi}_t^{NT}$ to ξ_{t0}, it is sufficient to establish the convergence of $\hat{G}_2^{NT}$ to the embedding from $\xi \in \Omega_\xi$ to Dirac_ξ, denoted by id_2.

To formalize this idea, we proceed in three steps. First, we construct the space $\hat{G}^{NT}$ belongs to. Define $\mathscr{P}(\Omega_{\alpha,\beta})$ and $\mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta}))$ as the set of probability distributions on $\Omega_{\alpha,\beta}$ and the set of continuous mappings from $\Omega_{\alpha,\beta}$ to $\mathscr{P}(\Omega_{\alpha,\beta})$, respectively. We can similarly define $\mathscr{P}(\Omega_\xi)$ and $\mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi))$. We define metrics on $\mathscr{P}(\Omega_{\alpha,\beta})$, $\mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta}))$, $\mathscr{P}(\Omega_\xi)$, and $\mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi))$. Because $\Omega_{\alpha,\beta}$ is separable, metrizable, and compact, then $\mathscr{P}(\Omega_{\alpha,\beta})$ is separable, metrizable and compact in the weak topology (Theorems 15.11 and 15.12 in Charalambos and Aliprantis (2013)):

$$G_{r1} \xrightarrow{w^*} G_1 \iff \int_{\Omega_{\alpha,\beta}} f(\alpha,\beta)(dG_{r1} - dG_1) \to 0, \ \forall\ f \in C_B(\Omega_{\alpha,\beta}),$$

where $C_B(\Omega_{\alpha,\beta})$ is the set of bounded continuous functions on $\Omega_{\alpha,\beta}$. Using Theorem 11.3.3 of Dudley (2018), the following metric $\|\cdot\|_P$ metrizes the w^*-topology:

$$\|G_1 - \tilde{G}_1\|_P := \sup\left\{\left|\int_{\Omega_{\alpha,\beta}} f(\alpha,\beta)d(G_1 - \tilde{G}_1)\right|,\ \|f\|_{BL} \leq 1\right\},$$

where f is bounded and Lipschitz, and $\|f\|_{BL} = \|f\|_L + \|f\|_\infty$ with

$$\begin{aligned}
\|f\|_L &:= \sup_{(\alpha,\beta)\neq(\tilde{\alpha},\tilde{\beta}),(\alpha,\beta),(\tilde{\alpha},\tilde{\beta})\in\Omega_{\alpha,\beta}} \left|f(\alpha,\beta) - f(\tilde{\alpha},\tilde{\beta})\right| / \|(\alpha-\tilde{\alpha},\beta-\tilde{\beta})\|, \\
\|f\|_\infty &:= \sup_{(\alpha,\beta)\in\Omega_{\alpha,\beta}} |f(\alpha,\beta)|,
\end{aligned}$$

where $\|\cdot\|$ refers to the Euclidean norm. As a result, $\mathscr{P}(\Omega_{\alpha,\beta})$ is compact in the

metric $\|\cdot\|_P$.

Denote by $G_1(\cdot|\alpha,\beta) \in \mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta}))$ a continuous mapping from $\Omega_{\alpha,\beta}$ to $\mathscr{P}(\Omega_{\alpha,\beta})$ and define the supremum metric as: for $G_1, \tilde{G}_1 \in \mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta}))$,

$$\|G_1 - \tilde{G}_1\|_1 = \sup_{(\alpha,\beta)\in\Omega_{\alpha,\beta}} \|G_1(\cdot|\alpha,\beta) - \tilde{G}_1(\cdot|\alpha,\beta)\|_P.$$

Similarly, define $G_2 \in \mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi))$,

$$\|G_2 - \tilde{G}_2\|_2 = \sup_{\xi\in\Omega_\xi} \|G_2(\cdot|\xi) - \tilde{G}_2(\cdot|\xi)\|_P.$$

Note that $\hat{G}_1^{NT}$ and $\hat{G}_2^{NT}$ in (9) are only defined at $(\alpha_{i0},\beta_{i0})_{i=1}^N$ and $(\xi_{t0})_{t=1}^T$, respectively. We now extend them to any point in $\Omega_{\alpha,\beta}$ and Ω_ξ so that they belong to $\mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta}))$ and $\mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi))$, respectively. For any N and T,

$$\begin{aligned}
\hat{G}_1^{NT}(a,b|\alpha,\beta) &:= \sum_{i\in\mathcal{S}_N(\alpha,\beta)} \frac{\prod_{r\neq i, r\in\mathcal{S}_N(\alpha,\beta)} \|(\alpha,\beta)-(\alpha_{r0},\beta_{r0})\|^2}{\sum_{i\in\mathcal{S}_N(\alpha,\beta)} \prod_{r\neq i: r\in\mathcal{S}_N(\alpha,\beta)} \|(\alpha,\beta)-(\alpha_{r0},\beta_{r0})\|^2} \mathrm{Dirac}_{(\hat{\alpha}_i^{NT},\hat{\beta}_i^{NT})}(a,b), \\
\hat{G}_2^{NT}(e|\xi) &:= \sum_{t\in\mathcal{S}_T(\xi)} \frac{\prod_{s\neq t: s\in\mathcal{S}_T(\xi)} |\xi-\xi_{s0}|^2}{\sum_{t\in\mathcal{S}_T(\xi)} \prod_{s\neq t: s\in\mathcal{S}_T(\xi)} |\xi-\xi_{s0}|^2} \mathrm{Dirac}_{\hat{\xi}_t^{NT}}(e),
\end{aligned} \tag{10}$$

where $\mathcal{S}_N(\alpha,\beta) = \left\{i : \|(\alpha_{i0},\beta_{i0})-(\alpha,\beta)\| \le \frac{\ln N}{\sqrt{N}}\right\}$ and $\mathcal{S}_T(\xi) = \{t : |\xi_{t0} - \xi| \le T^{-\frac{1}{4}}\}$. Under Assumption 4, both $\mathcal{S}_N(\alpha,\beta)$ and $\mathcal{S}_T(\xi)$ are asymptotically nonempty for any $(\alpha,\beta,\xi) \in \Omega_{\alpha,\beta}\times\Omega_\xi$ (see Remark 1 in Appendix D). Therefore, $\hat{G}^{NT} = (\hat{G}_1^{NT}, \hat{G}_2^{NT})$ is well-defined in $\Omega_{\alpha,\beta}\times\Omega_\xi$ when N and T are large enough. Moreover, for any $1\le i\le N$ and $1\le t\le T$, $\hat{G}^{NT}(a,b,e|\alpha_{i0},\beta_{i0},\xi_{t0}) = (\mathrm{Dirac}_{(\hat{\alpha}_i^{NT},\hat{\beta}_i^{NT})}(a,b), \mathrm{Dirac}_{\hat{\xi}_t^{NT}}(e))$.

Second, we rewrite the large-sample equivalence of $\mathscr{L}_{NT}(\theta^{NT})$ using (G,h) as in (9) and state the corresponding identification condition. For any $h \in \bar{\Theta}_h$, $G = (G_1,G_2) \in \mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta})) \times \mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi))$:

$$\begin{aligned}
\mathscr{L}^0(G,h;F_0) &= \int \mathbb{E}_0\left[\log\left(\frac{\exp\{h_Y(a+e+x'b)\}}{1+\sum_{y=1}^L \exp\{h_y(a+e+x'b)\}}\right)\Big|x,\alpha,\beta,\xi\right] dG_1(a,b;\alpha,\beta)dG_2(e;\xi)dF_0(x,\alpha,\beta,\xi) \\
&= \int \sum_{y=0}^L \left[\log\left(\frac{\exp\{h_y(a+e+x'b)\}}{1+\sum_{y=1}^L \exp\{h_y(a+e+x'b)\}}\right)\right] \frac{\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}{1+\sum_{y=1}^L \exp\{h_{y0}(\alpha+\xi+x'\beta)\}} dG_1(a,b;\alpha,\beta)dG_2(e;\xi)dF_0(x,\alpha,\beta,\xi),
\end{aligned}$$

where F_0 is the joint distribution of $(X_{it},\alpha_{i0},\beta_{i0},\xi_{t0})$. Note that by Gibbs' in-

equality, we obtain: for any $(\alpha', \beta', \xi', (h_y)_{y=1}^L)$,

$$\sum_{y=0}^{L}\left[\log\left(\frac{\exp\{h_y(\alpha'+\xi'+x'\beta')\}}{1+\sum_{y=1}^{L}\exp\{h_y(\alpha'+\xi'+x'\beta')\}}\right)\right]\frac{\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}{1+\sum_{y=1}^{L}\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}$$
$$\leq\sum_{y=0}^{L}\left[\log\left(\frac{\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}{1+\sum_{y=1}^{L}\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}\right)\right]\frac{\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}{1+\sum_{y=1}^{L}\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}.$$

$(\mathrm{id}_1, \mathrm{id}_2, (h_{l0})_{l=1}^L)$ is then a maximizer of $\mathscr{L}^0(G, (h_l)_{l=1}^L; F_0)$ in $\mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta})) \times \mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi)) \times \bar{\Theta}_h$. The next assumption states its uniqueness.

Assumption 6 (Identification)**.** *$\mathscr{L}^0(G, (h_l)_{l=1}^L; F_0)$ is uniquely maximized at (id_1, id_2, h_0) in $\mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta})) \times \mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi)) \times \bar{\Theta}_h$.*

In Appendix D.1, we show that Theorem 2 implies Assumption 6.
Lastly, we impose some regularity conditions on the log-likelihood function and $\mathscr{L}^0(G, h; F)$. Define

$$(\tilde{\alpha}, \tilde{\beta})\left((x_{it}, y_{it})_{t=1}^T; \{\xi_t\}_{t=1}^T, h\right) := \underset{(\alpha,\beta)\in\Omega_{\alpha,\beta}}{\arg\max}\frac{1}{T}\sum_{t=1}^{T}\log\frac{\exp\{h_{y_{it}}(\alpha+\xi_t+x_{it}'\beta)\}}{1+\sum_{y=1}^{L}\exp\{h_y(\alpha+\xi_t+x_{it}'\beta)\}}$$

i.e., the maximizer of the log-likelihood corresponding to individual i given $(y_{it}, x_{it}, \xi_t)_{t=1}^T$, and

$$\tilde{\xi}\left((x_{it}, y_{it})_{i=1}^N; \{\alpha_i, \beta_i\}_{i=1}^N, h\right) := \underset{\xi\in\Omega_\xi}{\arg\max}\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{\exp\{h_{y_{it}}(\alpha_i+\xi+x_{it}'\beta_i)\}}{1+\sum_{y=1}^{L}\exp\{h_y(\alpha_i+\xi+x_{it}'\beta_i)\}}\right]$$

as the maximizer of the log-likelihood corresponding to time period t given $(y_{it}, x_{it}, \alpha_{i0}, \beta_{i0})_{i=1}^N$.

Assumption 7 (Regularity conditions)**.**

(a) Uniform convergence:

$$\sup_{\theta_1^{NT}\in\Omega_{\alpha,\beta}^N\times\Omega_\xi^T, h\in\bar{\Theta}_h}\left|\mathscr{L}_{NT}(\theta_1^{NT}, h) - \mathscr{L}_{NT}^0(\theta_1^{NT}, h)\right| \xrightarrow{p} 0. \tag{11}$$

(b) Stochastic equicontinuity: For any $\varepsilon, \eta > 0$, there exists a random $\Delta_{NT}(\varepsilon, \eta) > 0$, positive constants $T_{\varepsilon,\eta}$, $N_{\varepsilon,\eta}$, and δ_ε such that for any $T > T_{\varepsilon,\eta}$, $N > N_{\varepsilon,\eta}$, we have: $\Pr(\Delta_{NT}(\varepsilon, \eta) > \varepsilon) < \eta$; for any (i, r) with

$\|(\alpha_{i0}, \beta_{i0}) - (\alpha_{r0}, \beta_{r0})\| < \delta_\varepsilon$,

$$\sup_{\xi_t \in \Omega_\xi, 1 \le t \le T; h \in \bar{\Theta}_h} \|(\tilde{\alpha}, \tilde{\beta}) \left((x_{it}, y_{it})_{t=1}^T; \{\xi_t\}_{t=1}^T, h\right) - (\tilde{\alpha}, \tilde{\beta}) \left((x_{rt}, y_{rt})_{t=1}^T; \{\xi_t\}_{t=1}^T, h\right) \| < \Delta_{NT}(\varepsilon, \eta);$$

for any (t, s) *with* $|\xi_{t0} - \xi_{s0}| < \delta_\varepsilon$,

$$\sup_{(\alpha_i, \beta_i) \in \Omega_{\alpha,\beta}, 1 \le i \le N; h \in \bar{\Theta}_h} |\tilde{\xi} \left((x_{it}, y_{it})_{i=1}^N; \{\alpha_i, \beta_i\}_{i=1}^N, h\right) - \tilde{\xi} \left((x_{is}, y_{is})_{i=1}^N; \{\alpha_i, \beta_i\}_{i=1}^N, h\right) | < \Delta_{NT}(\varepsilon, \eta).$$

(c)

$$\sup_{(G,h) \in \mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta})) \times \mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi)) \times \bar{\Theta}_h} \left| \mathscr{L}^0(G, h; F_0^{NT}) - \mathscr{L}^0(G, h; F_0) \right| \xrightarrow{p} 0. \tag{12}$$

where F_0^{NT} *is the empirical distribution of* $(x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0})_{1 \le i \le N; 1 \le t \le T}$.

Assumptions 7(a) and 7(c) extend the usual uniform convergence condition in the sieve setting (e.g., Condition 3.5 in Chen (2007)) to ours. Assumption 7(a) requires the uniformity over an increasing number of fixed-effects parameters and h; Assumption 7(c) imposes the uniformity over mapping G and h. The equicontinuity condition in Assumption 7(b) regularizes the large-sample dependence of individual-specific and time-specific maximum likelihood estimators, $(\tilde{\alpha}, \tilde{\beta}) \left((x_{it}, y_{it})_{t=1}^T; \{\xi_t\}_{t=1}^T, h\right)$ and $\tilde{\xi} \left((x_{it}, y_{it})_{i=1}^N; \{\alpha_i, \beta_i\}_{i=1}^N, h\right)$, on $(\alpha_{i0}, \beta_{i0})$ and ξ_{t0}, respectively. We will use this condition to construct a compact subset of $\mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta})) \times \mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi))$ that contains both $\hat{G}^{NT}$ and $(\mathrm{id}_1, \mathrm{id}_2)$ so that we can obtain the consistency in this compact subset. As we show in Appendix D.2, some restrictions on the link function, e.g., the (local) strong log-concavity, implies Assumption 7(b). In the same appendix, we show that Assumptions 7(a) and 7(c) can be achieved by additional regularity conditions on covariates X_{it}, $(\xi_t)_{t \ge 1}$, and the likelihood function.

We now state the consistency result. The proof is in Appendix D.

Theorem 3 (Consistency)**.** *Suppose that Assumptions 1, 2, 4–7 hold. Then,* $\|\hat{G}_1^{NT} - id_1\|_1 \xrightarrow{p} 0$, $\|\hat{G}_2^{NT} - id_2\|_2 \xrightarrow{p} 0$, *and* $\|\hat{h} - h_0\|_h \xrightarrow{p} 0$ *as* $N, T \to \infty$.

The convergences $\|\hat{G}_1^{NT} - \mathrm{id}_1\|_1 \to 0$ and $\|\hat{G}_2^{NT} - \mathrm{id}_2\|_2 \to 0$ in Theorem 3 imply $\sup_{i=1,\ldots,N} \left|(\hat{\alpha}_i^{NT}, \hat{\beta}_i^{NT}) - (\alpha_{i0}, \beta_{i0})\right| \xrightarrow{p} 0$ and $\sup_{t=1,\ldots,T} \left|\hat{\xi}_t^{NT} - \xi_{t0}\right| \xrightarrow{p} 0$. When the functions in $\bar{\Theta}_h$ are uniformly bounded and $\|\cdot\|_h$ is a stronger norm than the sup norm on $\bar{\Theta}_h$, $\|\hat{h} - h_0\|_h \to 0$ in Theorem 3 implies $\sup_{l=0,\ldots,L; v} |\hat{g}_l(v) - g_{l0}(v)| \to 0$, where

$\hat{g}_l = \exp\{\hat{h}_y\}/(1+\sum_{l=1}^{L}\exp\{\hat{h}_l\})$ is the plug-in estimator of g_{l0}. Consequently, we can consistently estimate the marginal effect $\frac{\partial g_0(y;x'\beta_{i0}+\alpha_{i0}+\xi_{t0})}{\partial x}$ and its average over time for individual i, denoted by

$$\text{AME}_i(y) = \mathbb{E}_{\xi,x|\alpha_{i0},\beta_{i0}}\left[\frac{\partial g_0(y;x'\beta_{i0}+\alpha_{i0}+\xi)}{\partial x}\right] \tag{13}$$

by plugging $(\hat{\alpha}_i^{NT},\hat{\beta}_i^{NT})_{i=1}^N$, $(\hat{\xi}_t^{NT})_{t=1}^T$, and $(\hat{g}_l)_{l=0}^L$ in the finite-sample analogue. In Section 4, we use Monte Carlo simulations to verify these implications.[8]

3.3 Implementation

Restrictions on the sieve space. Along the lines of Theorems 1 and 2 and Assumption 6 (identification condition), one may need to impose some restrictions on h to guarantee the consistency, e.g., the monotonicity in Assumption 1(b). As a result, the sieve space Θ_h^{NT} used in (8) should correspondingly incorporate such restrictions. For instance, suppose the model is binary and we use the polynomial sieve of order d to estimate h_1 in (8):

$$\left\{\sum_{r=0}^{d} a_r v^r : a_r \in \mathcal{A},\ r=0,\ldots,d\right\}.$$

where $\mathcal{A}$ is a compact subset of $\mathbb{R}$. Denote by $\mathcal{D}$ the range of the index. That the link function $g(1;v)$ (or equivalently $h_1(v)$) is increasing in v amounts to imposing the following linear inequalities:[9] for any $v \in \mathcal{D}$,

$$\sum_{r=1}^{d} a_r r v^{r-1} > 0.$$

Another example is log-concavity of $g(1;v) = \exp\{h_1(v)\}/(1+\exp\{h_1(v)\})$. This concavity condition amounts to imposing the following nonlinear inequalities: for

[8] As previously discussed, we can only identify (and consistently estimate) (β_{i0},α_{i0}) and ξ_{t0} up to a shift and scale unless we normalize the $(\beta_i^{(k)},\alpha_{i'},\xi_t)$ to their true values for some i, i', t, and k. In Appendix G, we report the estimates for (β_{i0},α_{i0}) and ξ_{t0} under such normalizations to provide further support for Theorem 3.

[9] An alternative way to incorporate the monotonicity constraint is to first estimate h without such constraints and monotonize the estimated link function. See Chernozhukov et al. (2009) for an example.

any $v \in \mathcal{D}$,

$$\left(\sum_{r=2}^{d} a_r r(r-1)v^{r-2}\right)\left(1+\exp\left\{\sum_{r=0}^{d} a_r v^r\right\}\right) - \left(\sum_{r=1}^{d} a_r r v^{r-1}\right)^2 \exp\left\{\sum_{r=0}^{d} a_r v^r\right\} < 0.$$

In practice, one can use a grid of values of $v \in \mathcal{D}$ to implement these inequalities as a set of constraints on sieve coefficients $(a_r)_{r=0}^d$.

Normalization of $\beta_{i^\star}^{(1)}$. As discussed below Theorem 2, we can identify the sign of $\beta_{i^\star}^{(1)}$ and accordingly normalize it to 1 or -1 in estimation. In some settings, the researcher can derive the sign from economic theory or past research (e.g., negative price slope due to the law of demand). Without any prior on the sign of $\beta_{i^\star}^{(1)}$, one can practically implement the sieve MLE with the two normalizations and choose the estimates that generate a higher likelihood.

4 Monte Carlo Simulations

In this section, we first use Monte Carlo simulations to illustrate the consequences of ignoring slope heterogeneity. Second, we investigate the finite-sample performance of the sieve maximum likelihood estimator (8) and verify our consistency results. For both tasks, we consider a static binary choice model that mimics the establishment of exportation/importation in trade: for $1 \le i < j \le N$,

$$\Pr(y_{ij}=1|w_{ij},\beta_i,\alpha_i^{(1)},\alpha_j^{(2)}) = g_0(w_{ij}^{(1)}\beta^{(1)} + w_{ij}^{(2)}\beta_i^{(2)} + \alpha_i^{(1)} + \alpha_j^{(2)}),$$

where $\alpha_i^{(1)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}[-1,1]$, $\alpha_j^{(2)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}[-1,1]$. Individual-specific slopes $\beta_i^{(2)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0.2,1.2]$. Moreover, $w_{ij}^{(1)} = 0.5\alpha_i^{(1)} + 0.2\alpha_j^{(2)} + \mu_{ij}^{(1)}$ with $\mu_{ij}^{(1)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}[-1,1]$ and $w_{ij}^{(2)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}[-1,0]$. We set $\beta^{(1)} = 1$. $(\alpha_i^{(1)})_{i=1}^N$, $(\alpha_i^{(2)})_{i=1}^N$, $(\mu_{ij})_{i,j}$, and $(w_{ij}^{(2)})_{i,j}$ are independent. This data generating process and the sampling procedure satisfy Assumptions 1 and 2. In Appendix G, we report the Monte Carlo results using another data generating process that violates Assumption 1.

Consequence of ignoring slope heterogeneity. Suppose that g_0 is a logit (probit) link function. For each sample size $N \in \{50,\ 100,\ 200\}$, we generate $R = 200$ sets of outcomes $(y_{ij})_{1\le i<j\le N}$. For each replication, we estimate a logit (probit) model that restricts $\beta_i^{(2)} = \beta^{(2)}$ and one without such restrictions. The former logit (probit) MLE ignores the heterogeneity in the true $\beta_i^{(2)}$; the latter one

Table 1: Consequence of ignoring slope heterogeneity: Logit and probit cases

Scenario $g_0(\delta)$	$\frac{\exp\{\delta\}}{1+\exp\{\delta\}}$		$\Phi(\delta)$	
	$\text{AME}_i^{(1)}$	$\text{AME}_i^{(2)}$	$\text{AME}_i^{(1)}$	$\text{AME}_i^{(2)}$
$N = 50$, MLE with $\beta_i^{(1)} = \beta^{(1)}$	0.2395	0.1928	0.1643	0.1517
MLE with $\beta_i^{(1)}$	0.0334	0.2018	0.0401	0.1949
$N = 100$, MLE with $\beta_i^{(1)} = \beta^{(1)}$	0.2702	0.1895	0.2154	0.1675
MLE with $\beta_i^{(1)}$	0.0177	0.1552	0.0221	0.1417
$N = 200$, MLE with $\beta_i^{(1)} = \beta^{(1)}$	0.2191	0.1664	0.1536	0.1299
MLE with $\beta_i^{(1)}$	0.0115	0.1072	0.0152	0.0937

Notes: We estimate the logit and probit models by normalizing $\alpha_1^{(1)}$ to its true value $\alpha_{10}^{(1)}$. Each cell corresponds to the average distance metrics between the estimated object and its true value over 200 repetitions for a given sample size N, scenario of true link function g_0, and the model used in the MLE (logit, probit). For $\text{AME}_i^{(k)}$ with $k = 1, 2$, the distance metrics is defined as $\sqrt{\sum_{i=1}^N (\widehat{\text{AME}}_i^{(k)} - \text{AME}_{i0}^{(k)})^2/N}$ where 0 refers to the true values.

is a special case of Theorem 3 in which the link function is known. For each model, we compare the distributions of estimated average marginal effects of $w_{ij}^{(1)}$ and of $w_{ij}^{(2)}$ across $i = 1, \ldots, N$ (referred to as $\text{AME}_i^{(1)}$ and $\text{AME}_i^{(2)}$ in (13), respectively) to the true ones by using distance metrics defined as $\sqrt{\frac{\sum_{i=1}^N (\widehat{\text{APE}}_i^{(k)} - \text{APE}_{i0}^{(k)})^2}{N}}$ for $k = 1, 2$ where 0 refers to the true values. For each object of interest, we report the average distance over 200 repetitions.

Table 1 summarizes our results. First, note that the distance metrics for AMEs based on the MLE without the restrictions $\beta_i^{(2)} = \beta^{(2)}$ (second row in each panel) decreases as the sample size increase, which aligns with the consistency result in Theorem 3. Compared to the MLE that restricts $\beta_i^{(2)} = \beta^{(2)}$ (the first row in each panel), the AMEs of $w_{ij}^{(2)}$ predicted by the MLE without the homogeneity restrictions are more precise when $N = 100$ and 200. Besides, even though only imposed on the slope of $w_{ij}^{(2)}$, the homogeneity restrictions $\beta_i^{(2)} = \beta^{(2)}$ also deteriorate the precision of the predicted $\text{AME}_i^{(1)}$. In Section 5, we will further investigate the extent to which the slope homogeneity restrictions may bias the analysis of predicted AMEs using real data.

Finite-sample performance of the sieve MLE. We consider three scenarios of link functions: $g_0(\delta) = \frac{\exp\{\delta\}}{1+\exp\{\delta\}}$ (logit), $g_0(\delta) = \Phi(\delta)$ (probit), and $g_0(\delta) = \frac{\exp\{2\exp\{\delta\}\}}{1+\exp\{2\exp\{\delta\}\}}$. For each sample size $N \in \{50,\ 100,\ 200\}$ and each sce-

Table 2: Finite-sample performances: Polynomial sieves

Scenario $g_0(\delta) =$	$\frac{\exp\{\delta\}}{1+\exp\{\delta\}}$			$\Phi(\delta)$			$\frac{\exp\{2\exp\{\delta\}\}}{1+\exp\{2\exp\{\delta\}\}}$		
	$\text{AME}_i^{(1)}$	$\text{AME}_i^{(2)}$	g	$\text{AME}_i^{(1)}$	$\text{AME}_i^{(2)}$	g	$\text{AME}_i^{(1)}$	$\text{AME}_i^{(2)}$	g
$N = 50$, Logit	0.0334	0.2018	×	0.0403	0.1707	×	0.0565	0.1696	×
Probit	0.0327	0.2178	×	0.0401	0.1949	×	0.0578	0.1905	×
Poly. sieve, $d = 1$	0.0315	0.2071	0.0677	0.0412	0.1957	0.0571	0.0665	0.1825	0.1821
$d = 2$	0.0375	0.2073	0.0743	0.0419	0.1961	0.0603	0.0497	0.1480	0.1207
$d = 3$	0.0479	0.2261	0.0997	0.0509	0.2188	0.0802	0.0481	0.1508	0.1167
$d = 4$	0.0553	0.2282	0.1199	0.0571	0.2253	0.0926	0.0515	0.1519	0.1157
$N = 100$, Logit	0.0177	0.1552	×	0.0234	0.1352	×	0.0473	0.1197	×
Probit	0.0169	0.1573	×	0.0221	0.1417	×	0.0469	0.1252	×
Poly. sieve, $d = 1$	0.0171	0.1554	0.0646	0.0239	0.1420	0.0620	0.0520	0.1241	0.1483
$d = 2$	0.0184	0.1555	0.0663	0.0242	0.1420	0.0615	0.0299	0.1072	0.1247
$d = 3$	0.0207	0.1559	0.0665	0.0227	0.1421	0.0614	0.0245	0.1101	0.0890
$d = 4$	0.0285	0.1598	0.0892	0.0234	0.1423	0.0627	0.0247	0.1101	0.0865
$N = 200$, Logit	0.0115	0.1072	×	0.0166	0.0925	×	0.0483	0.0953	×
Probit	0.0112	0.1074	×	0.0152	0.0937	×	0.0462	0.0957	×
Poly. sieve, $d = 1$	0.0115	0.1072	0.0359	0.0167	0.0940	0.0299	0.0495	0.0962	0.1359
$d = 2$	0.0118	0.1072	0.0362	0.0161	0.0939	0.0303	0.0284	0.0806	0.1417
$d = 3$	0.0121	0.1072	0.0374	0.0153	0.0938	0.0287	0.0145	0.0845	0.0616
$d = 4$	0.0122	0.1074	0.0400	0.0153	0.0938	0.0285	0.0146	0.0845	0.0589

Notes: Each cell corresponds to the average distance metrics between the estimated object and its true value over 200 repetitions for a given sample size N, scenario of true link function g_0, and the model used in the MLE (logit, probit, or polynomial sieves of degree $d = 1, \ldots, 4$). For $\text{AME}_i^{(k)}$ with $k = 1, 2$, the distance metrics is defined as $\sqrt{\sum_{i=1}^N (\widehat{\text{AME}}_i^{(k)} - \text{AME}_{i0}^{(k)})^2/N}$ where 0 refers to the true values. For the sieve MLE, the distance corresponding to the link function is defined as $\sqrt{\sum_{m=1}^M \left(\hat{g}(\delta_m) - g_0(\delta_m)\right)^2/M}$ where $(\delta_m)_{m=1}^M$ is an equal-spaced (by 0.1) sequence of values covering the true range of the index in the data generating process.

nario, we generate $R = 200$ sets of outcomes $(y_{ij})_{1\le i<j\le N}$. For each replication, we implement the MLE using a logit model, a probit model, and polynomial sieves for function $h(\delta) = \ln\left(g(\delta)/(1-g(\delta))\right)$ with the sieve space being of order $d = 1$ to $d = 4$, respectively. For the logit and probit MLE, we normalize $\alpha_1^{(1)}$ to its true value $\alpha_{10}^{(1)}$. For the sieve MLE, we further normalize $\beta^{(1)} = 1$ and $\alpha_1^{(2)}$ to its true value $\alpha_{10}^{(2)}$. For each model, we compare the estimated distributions of $\text{AME}_i^{(1)}$ and $\text{AME}_i^{(2)}$ across $i = 1, \ldots, N$ by using distance metrics $\sqrt{\frac{\sum_{i=1}^N (\widehat{\text{AME}}_i^{(k)} - \text{AME}_{i0}^{(k)})^2}{N}}$ for $k = 1, 2$. For the sieve MLE, we also compare the estimated link function to the true one and by the distance metric $\sqrt{\frac{\sum_{m=1}^M (\hat{g}(\delta_m) - g_0(\delta_m))^2}{M}}$ where $(\delta_m)_{m=1}^M$ is an equal-spaced (by 0.1) sequence of values covering the range of the index in the data generating process. For each object of interest, we report the average distance over 200 repetitions.

Our main results are summarized in Table 2. In Appendix G, we report the distance statistics for other objects ($\beta_i^{(1)}$, $\alpha_i^{(1)}$, and $\alpha_i^{(2)}$). First, in the logit scenario, because the polynomial sieve space contains the true link function, the distance metric for the link function corresponding to polynomial sieve MLE (columns

"g") decreases as the sample size increases, suggesting a convergence towards the true link function. In other scenarios, the similar pattern holds. Due to the improved link function estimation, the distances between the distributions of AMEs predicted by the sieve MLE and the true ones decrease as N increases. When $N = 50$, these distance metrics corresponding to the sieve MLE are greater than those for the correctly specified models (i.e., the row "Logit" in the logit scenario and the row "Probit" in the probit one). As N increases, the sieve MLE performs better, achieving similar precision to the correctly specified models when $N = 100$ and 200.

When the true link function is $\frac{\exp\{2\exp\{\delta\}\}}{1+\exp\{2\exp\{\delta\}\}}$, both the probit and logit MLEs are misspecified. For all the three sample sizes, the MLE with a polynomial sieve space of dimension $d \geq 2$ outperforms logit and probit models when predicting the distributions of $\text{AME}_i^{(1)}$ and $\text{AME}_i^{(2)}$. These results again confirm the consistency of the sieve MLE in Theorem 3 and its ability in attenuating the misspecification errors of the link function in finite sample.

5 Empirical Illustration

In this section, we revisit some empirical study in Helpman et al. (2008). In the original paper, the authors estimate trade flows and explicitly take into account firm selection into the export market. They first estimate the probability of establishing exportation from one country to another using a binary model. One can then control for the fraction of firms that export (consistently estimated from the first step) and the selection effect due to zero trade flows when estimating the gravity equation in the second step. In the empirical application, the first step is implemented as (see their equation 12 on page 455):

$$\Pr(T_{ij} = 1 \mid \text{dist}_{ij}, w_{ij}, \zeta_i, \xi_j) = \Phi\left(-\gamma \text{dist}_{ij} + w_{ij}'\kappa + \zeta_i + \xi_j\right), i, j = 1, \ldots, N, i \neq j, \quad (14)$$

where $T_{ij} = 1$ when country j exports to i and zero otherwise, dist_{ij} is the log distance between i and j, w_{ij} is a vector of observed country-pair specific variables, ζ_i (ξ_j) is an importer (exporter) fixed effect, and Φ is the standard normal cumulative distribution function. Parameter γ is interpreted as the constant marginal effect of log distance on the probability of country j exporting to i.

Different from the original empirical setting, we allow γ to be country-specific

and the link function to be unknown:[10]

$$\Pr(T_{ij} = 1 \mid \text{dist}_{ij}, w_{ij}, \zeta_i, \xi_j) = g(-\gamma_j^{\exp}\text{dist}_{ij} + w_{ij}'\kappa + \zeta_i + \xi_j), i, j = 1, \ldots, N, i \neq j. \quad (15)$$

Specification (15) relaxes the restriction of constant marginal effect of log distance in two ways. First, the same country i may react differently when importing from different countries j and j' of similar distances if $\gamma_j^{\exp} \neq \gamma_{j'}^{\exp}$. Second, two countries i and j can have different distance effects when exporting to the other if $\gamma_i^{\exp} \neq \gamma_j^{\exp}$. Besides, treating the link function as parameters to be estimated can attenuate potential bias due to imposing a known link function, e.g., the thin tail of Gaussian distribution.

We estimate (15) using the 1986 worldwide trade data sample of Helpman et al. (2008) and compare several objects of interest across different models to shed light on the consequences of imposing Gaussian link function and slope homogeneity. The data include $N = 158$ countries. As in the original paper, we remove Congo as an exporter from the sample because it did not export to anyone in 1986.[11] This leaves us with $24,649$ observations of directed trade flows (exportation) . We use the set of controls in the second column of Table 1 of Helpman et al. (2008) as w_{ij} in (15). Most covariates in w_{ij} are discrete (e.g., whether two countries have common border) with the exception of a continuous measure of common religion belief between two countries. To apply our identification argument, one can choose dist_{ij} as the compensating variable as it is continuous and has a relatively large range (between -0.151 and 5.661). Such a choice is compatible with Assumption 2(e) when the distribution of the importing-country fixed effect ζ_i is independent of dist_{ij} conditional on w_{ij} (see the discussion below Assumption 1(e)).

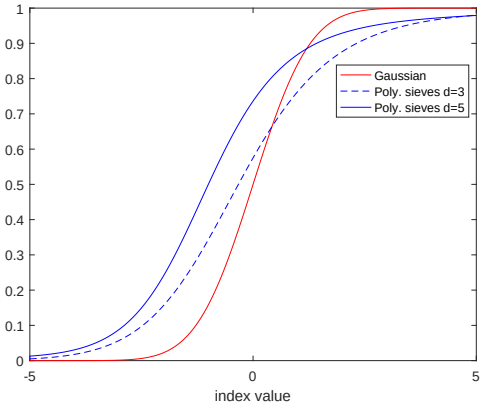
In Figure 1(a), we compare the Gaussian cumulative probability function (red; the link function in probit model) to the estimated link function g in models (15). Function $h(\cdot) = \ln(g(\cdot)/(1 - g(\cdot)))$ in the latter models is estimated by polynomial sieves of degree $d \in \{3, 5\}$ (blue). We find that the estimated link functions have thicker left and right tails than Gaussian distribution. As the sieve dimension increases, the right tail of the estimated link function approaches the Gaussian one; however, the estimated left tail seems to be still thicker than the

[10] Specification (15) can be obtained by relaxing some functional-form restrictions in Helpman et al. (2008). For example, one can relax $\tau_{ij}^{\varepsilon-1} = D_{ij}^{\gamma} e^{-u_{ij}}$ (on page 453 in the original paper) to $\tau_{ij}^{\varepsilon-1} = D_{ij}^{\gamma_j} e^{-u_{ij}}$ where D_{ij} is the distance between i and j and u_{ij} is an unmeasured trade friction.

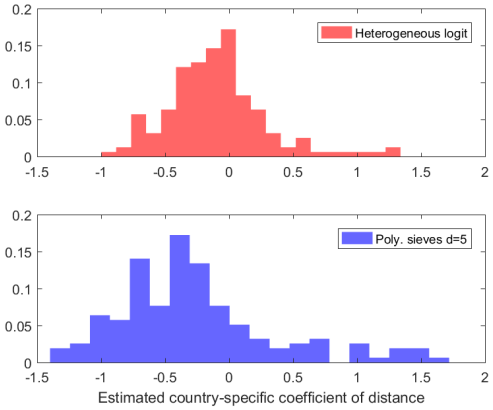
[11] See footnote 23 on page 459 of Helpman et al. (2008) .

Figure 1: ESTIMATED LINK FUNCTION AND $-\gamma_j^{\text{EXP}}$

(a) Estimated link function



(b) Estimated $-\gamma_j^{\text{exp}}$



Gaussian distribution. This difference in the link function's left tail may affect the slope estimates. Intuitively, to fit the same observed link probability in the data, Gaussian link function may allocate a greater index value than the link function estimated by polynomial sieves with $d = 5$. Given a covariate value (say, distance), this will lead to an upward bias in the estimate of the corresponding coefficient. The comparisons of the estimated $-\gamma_j^{\text{exp}}$ in Figure 1(b) align with this intuition. We find that the distribution of $-\gamma_j^{\text{exp}}$ estimated by the probit model (red histogram) is overall shifted towards the right relative to that corresponding to the sieve estimation with $d = 5$ (blue histogram), with the estimated mean of $-\gamma_j^{\text{exp}}$ being -0.0952 versus -0.2578.

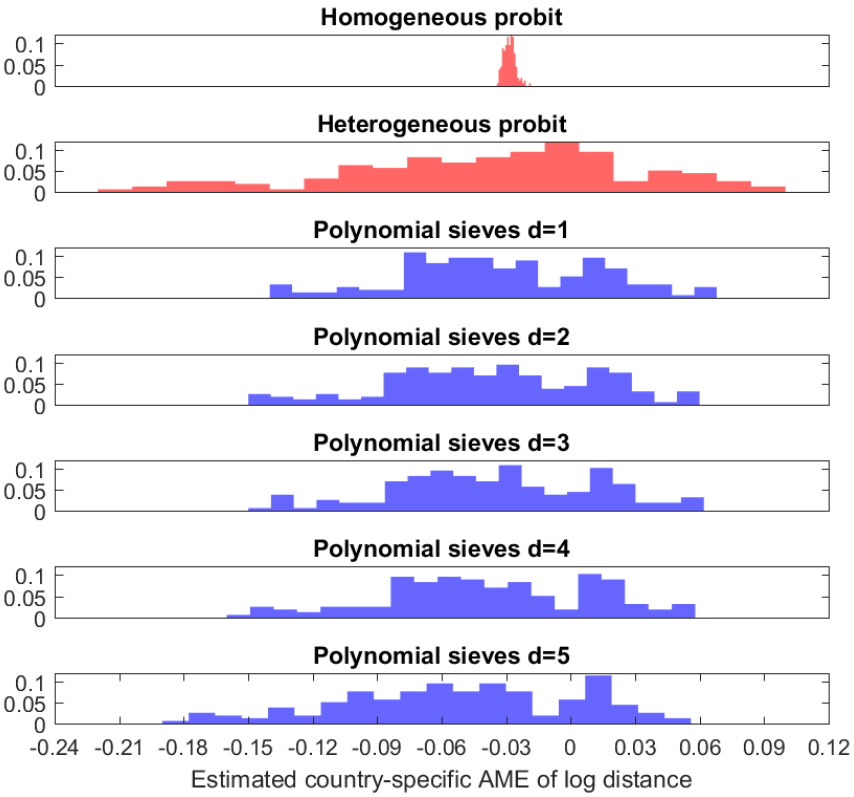
We now turn to exporting-country-specific marginal effects of dist_{ij} defined as

$$\text{AME}_j^{\text{dist}} = -\frac{1}{N}\sum_{i=1}^{N} \gamma_j^{\text{exp}} \partial_\delta g(-\gamma_j^{\text{exp}} \text{dist}_{ij} + w_{ij}'\kappa + \zeta_i + \xi_j),$$

where N is the total number of countries. In Figure 2, each subfigure illustrates the distribution of $\text{AME}_j^{\text{dist}}$ estimated by a probit model (14) with homogeneous γ (first red histogram), a probit model (15) with heterogeneous γ_j^{exp} (second red histogram), and models (15) with heterogeneous γ_j^{exp} and $h(\delta) = \ln(g(\delta)/(1-g(\delta)))$ being estimated by polynomial sieves of degree $d = 1, \ldots, 5$ (blue histograms).

In the homogeneous probit model, the restriction $\gamma_j^{\text{exp}} = \gamma$ assumes away any variation in γ_j^{exp}. As a result, $\text{AME}_j^{\text{dist}}$ varies across exporting countries solely due

Figure 2: Estimated distribution of $\text{AME}_j^{\text{DIST}}$ across countries



to the variation in $\frac{1}{N}\sum_{i=1}^{N}\phi(-\gamma \text{dist}_{ij} + w_{ij}'\kappa + \zeta_i + \xi_j)$ where ϕ is the standard normal density. In the presence of potential heterogeneity in γ_j^{exp}, this restricted model may mechanically understate the dispersion of $\text{AME}_j^{\text{dist}}$. Our findings in the red histograms align with this intuition. Both probit models with and without the restriction $\gamma_j^{\text{exp}} = \gamma$ predict similar distribution means (-0.0285 and -0.0353, respectively). However, the standard deviation of $\text{AME}_j^{\text{dist}}$ estimated by the homogeneous probit model is more than 20 times smaller than that estimated by the probit model with heterogeneous γ_j^{exp} (0.0027 vs 0.0662). Even if we relax the probit restriction and flexibly estimate the link function, as illustrated by the blue histograms in Figure 2, we still find substantial dispersion in the estimated $\text{AME}_j^{\text{dist}}$; the standard deviation range from 0.0458 to 0.0545 in these scenarios, smaller than (but close to) the prediction by the heterogeneous probit model.

Furthermore, we decompose the dispersion in the estimated $\text{AME}_j^{\text{dist}}$ into two components: the variance explained by $-\gamma_j^{\text{exp}}$ and that explained by

$\frac{1}{N}\sum_{i=1}^{N}\partial_\delta g(-\gamma_j^{\exp}\text{dist}_{ij}+w_{ij}'\kappa+\zeta_i+\xi_j)$. The results are summarized in Table 3. We find that $-\gamma_j^{\exp}$ explains most of the variation in $\text{AME}_j^{\text{dist}}$ (around 94%) across all models with heterogeneous $\gamma_j^{\exp}$. Combined with Figure 2, these results suggest that imposing homogeneity on $\gamma_j^{\exp}$ across countries may substantially restrict the variation in $\text{AME}_j^{\text{dist}}$ and distort the analysis of marginal effect of distance on establishing exportation/importation between country pairs.

Table 3: VARIANCE DECOMPOSITION OF $\text{AME}_j^{\text{DIST}}$

	Variance	Variance explain by $-\gamma_j^{\exp}$	
	$\text{Var}\left(\text{AME}_j^{\text{dist}}\right)$	$\text{Var}\left(\mathbb{E}\left[\text{AME}_j^{\text{dist}}\middle\vert\gamma_j^{\text{dist}}\right]\right)$	in %
Probit with $\gamma_j^{\exp}$	0.0658	0.0619	94.12%
Polynomial sieves, $d=1$	0.0458	0.0375	93.21%
$d=2$	0.0476	0.0444	93.29%
$d=3$	0.0470	0.0440	93.69%
$d=4$	0.0478	0.0446	93.47%
$d=5$	0.0542	0.0507	93.49%

Notes: We obtain $\mathbb{E}\left[\text{AME}_j^{\text{dist}}\middle\vert\gamma_j^{\text{dist}}\right]$ by regressing $\text{AME}_j^{\text{dist}}$ on γ_j^{dist} and its polynomial terms (up to order 5).

6 Conclusion

In this paper, we study a class of two-way fixed effects index function models with a nonparametric link function and individual- (or time-) specific slopes in index. This relaxes the practice of specifying a known link function and the homogeneity restriction on covariate slopes, both of which are commonly adopted in applied and related econometric works. We show the identification of the fixed effects parameters and the link function when both N and T are large. We also propose a nonparametric consistency result for the fixed effects sieve maximum likelihood estimators. We revisit the study of establishing between-country exportation in Helpman et al. (2008) and illustrate the consequences of imposing Gaussian link function and homogeneity on the slope of distance.

Our identification and consistency results provide a foundation for inference. Existing results in the literature of two-way fixed effects panel models with large N and large T mostly deal with inferences when the link function is known. Many focus on correcting the asymptotic bias in the estimates of objects of interest

such as the slopes and average marginal effects. The link function in our setting is nonparametrically specified and estimated. Besides, the number of individual (or time)-specific slopes increases asymptotically. Both complicate the inference, leaving the applicability of the existing bias corrections to our setting an open question. In a recent work, Jochmans and Weidner (2024) propose an inference framework in which one could view the estimated slopes (and individual- or time-fixed effects) as noisy measurements around the true ones. A key assumption in their approach is that the noisy measurements are i.i.d. and shrink to zero with a uniform parametric rate. One interesting avenue of future research is to adapt their framework to the case of nonparametric link function.

Appendix

Notation and some useful lemmas. For any $p \geq 1$ and any two vectors x and y in $\mathbf{R}^p$, we let $\langle x, y\rangle$ denote the usual Euclidean inner product of x with y. Thus, the Euclidean norm is given by $\|x\| = \sqrt{\langle x, x\rangle}$. For a matrix A, we denote A' as the transpose of A. For a real symetric matrix $A \in \mathbf{R}^{n\times n}$, we let $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ denote its real eigenvalues. For any real matrix $A \in \mathbf{R}^{n\times m}$, $\|A\|_2 := \sqrt{\lambda_1(A'A)}$ denotes the spectral norm (i.e., the operator norm induced by the Euclidean norm), $\|A\|_{\mathrm{F}} := \sqrt{\mathrm{tr}(A'A)}$ denotes the Frobenius norm, and $\|A\|_{\max} := \max_{i=1,\ldots,n;j=1,\ldots,m} |A_{ij}|$ denotes the element-wise max norm. We use a.s. to refer to "almost surely" and a.e. "almost everywhere in the domain".
Define, for all i such that $\sum_{t=1}^{\infty} M_{it}^{NT} \geq 1$, for all $(y, x) \in \mathcal{Y} \times \mathcal{X}_i$,

$$\widehat{\Gamma}_i(y, x) \equiv \frac{\frac{1}{h_T \sum_{t=1}^{\infty} M_{it}^{NT}} \sum_{t=1}^{\infty} M_{it}^{NT} K\left(\frac{X_{it}-x}{h_T}\right) \mathbf{1}\{Y_{it} = y\}}{\frac{1}{h_T \sum_{t=1}^{\infty} M_{it}^{NT}} \sum_{t=1}^{\infty} M_{it}^{NT} K\left(\frac{X_{it}-x}{h_T}\right)}.$$

The following lemmas will be used in the proofs of Theorems 1 and 2. Their proofs can be found in Online Appendix F.

Lemma 1. *Suppose that Assumption 1 holds. Moreover, $K : \mathbf{R}^K \to \mathbf{R}$ is bounded, $\int |K|^{2+\delta}(u)du < \infty$ for some $\delta > 2/(\mu - 1) > 0$, and $h_T \to 0$ and $Th_T \to \infty$. Then, as N, T tend to infinity, conditional on $\mathcal{F}$,*

$$\sup_{i, x\in\mathcal{X}_i} \left|\widehat{\Gamma}_i(\overline{y}, x) - \Gamma_i(\overline{y}, x)\right| = o_p(1),$$

where $\Gamma_i(\overline{y}, x) \equiv c_{g(\overline{y};x'\beta_i+\alpha_i+\cdot),x^{(2)}}$ *in Assumption 1(e). Moreover,* $\Gamma_i(\overline{y}, x)$ *is strictly monotonic in* $x'\beta_i + \alpha_i$.

Due to Lemma 1, the econometrician knows the true value of $\Gamma_i(\bar{y}; x)$ for any $x \in \mathcal{X}_i$ and i when T and N are infinity.

Lemma 2. $i \longleftrightarrow i'$ *in the compensating network* $\mathcal{G}^\infty$ *if and only if there exist* $(\tilde{x}^{(1)k}, \tilde{x}^{(2)k}) \in \mathcal{X}_i$, $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_{i'}$ *for* $k = 1, 2, 3$ *such that*

$$\Gamma_i(\overline{y}, (\tilde{x}^{(1)k}, \tilde{x}^{(2)k})) = \Gamma_{i'}(\overline{y}, (x^{(1)k}, x^{(2)k})),$$

with

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix} \text{ or } \begin{bmatrix} 1 & \tilde{x}^{(1)1} & \tilde{x}^{(2)1} \\ 1 & \tilde{x}^{(1)2} & \tilde{x}^{(2)2} \\ 1 & \tilde{x}^{(1)3} & \tilde{x}^{(2)3} \end{bmatrix}$$

being nonsingular.

A Proof of Theorem 1

Recall that $\{\mathcal{I}_r : r = 1, 2, \ldots\}$ is the partition of $\mathbb{N} \setminus \mathcal{I}_0$ induced by $\mathcal{G}^\infty$.

Identification of $\mathcal{I}_0$ and $\mathcal{G}^\infty$. For any i, we can identify $\Gamma_i(\overline{y}, x)$ for $x \in \mathcal{X}_i$ by Lemma 1. Because Γ_i is strictly monotonic in $x'\beta_i + \alpha_i$, then $\beta_i^{(1)} = 0$ if and only if Γ_i does not depend on $x^{(1)} \in \mathcal{X}_i$. Since the latter is identified, we can then identify if $\beta_i^{(1)} = 0$ and $\mathcal{I}_0$. Using similar arguments, we can identify if $i \longleftrightarrow i'$ and therefore $\mathcal{G}^\infty$.

Suppose $i \to i'$ and $i, i' \in \mathcal{I}_r$ for some $r \geq 1$. Then, $\beta_i^{(1)} \neq 0$ and using Lemma 2, we can find $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_{i'}$ and $(\tilde{x}^{(1)k}, x^{(2)k}) \in \mathcal{X}_i$ with $\tilde{x}^{(1)k} = z_{i\to i'}(x^{(1)k}; x^{(2)k})$ for $k = 1, 2, 3$ such that the matrix

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(1)1} \\ 1 & x^{(1)2} & x^{(1)2} \\ 1 & x^{(1)3} & x^{(1)3} \end{bmatrix}$$

is nonsingular. Because of the strict monotonicity in Lemma 1, we have:

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(1)1} \\ 1 & x^{(1)2} & x^{(1)2} \\ 1 & x^{(1)3} & x^{(1)3} \end{bmatrix} \begin{pmatrix} (\alpha_{i'} - \alpha_i)/\beta_i^{(1)} \\ \beta_{i'}^{(1)}/\beta_i^{(1)} \\ (\beta_{i'}^{(2)} - \beta_i^{(2)})/\beta_i^{(1)} \end{pmatrix} = \begin{pmatrix} \tilde{x}^{(1)1} \\ \tilde{x}^{(1)2} \\ \tilde{x}^{(1)3} \end{pmatrix},$$

and solving this linear system identifies $(\alpha_{i'} - \alpha_i)/\beta_i^{(1)}$, $\beta_{i'}^{(1)}/\beta_i^{(1)}$ and $(\beta_{i'}^{(2)} - \beta_i^{(2)})/\beta_i^{(1)}$.

Now for a sequence $(i_1, i_2), \ldots, (i_{l-1}, i_l)$ with $i_k \in \mathcal{I}_r$ for $k = 1, \ldots, l$, by applying the arguments above, we can identify either $(\alpha_{i_{k-1}} - \alpha_{i_k})/\beta_{i_k}^{(1)}$, $\beta_{i_{k-1}}^{(1)}/\beta_{i_k}^{(1)}$ and $(\beta_{i_{k-1}}^{(2)} - \beta_{i_k}^{(2)})/\beta_{i_k}^{(1)}$, or $(\alpha_{i_k} - \alpha_{i_{k-1}})/\beta_{i_{k-1}}^{(1)}$, $\beta_{i_k}^{(1)}/\beta_{i_{k-1}}^{(1)}$ and $(\beta_{i_k}^{(2)} - \beta_{i_{k-1}}^{(2)})/\beta_{i_{k-1}}^{(1)}$ for $k = 2, \ldots, l$. Then, we identify $\beta_{i_k}^{(1)}/\beta_{i_1}^{(1)}$ for $k = 2, \ldots, l$, and therefore $(\alpha_{i_{k-1}} - \alpha_{i_k})/\beta_{i_1}^{(1)}$ and $(\beta_{i_{k-1}}^{(2)} - \beta_{i_k}^{(2)})/\beta_{i_1}^{(1)}$. This implies the identification of $(\alpha_{i_l} - \alpha_{i_1})/\beta_{i_1}^{(1)}$, $\beta_{i_l}^{(1)}/\beta_{i_1}^{(1)}$, and $(\beta_{i_l}^{(2)} - \beta_{i_1}^{(2)})/\beta_{i_1}^{(1)}$. The proof is completed.

B Proof of Theorem 2

The first statement. The identification of $(\alpha_i, \beta_i^{(1)})$ in the first statement follows from Theorem 1 and Assumption 3(a). We now use Assumption 3(b) to identify $\beta_{i^\star}^{(2)}$, which will imply the identification of $\beta_i^{(2)}$ for all i given the identification of $(\beta_i^{(2)} - \beta_{i^\star}^{(2)})/\beta_{i^\star}^{(1)}$ in Theorem 1.

Because of the definition of compensating variable in (5) and Theorem 1, we identify $z_{i^\star \to i}(x_{it}^{(1)}, x_{it}^{(2)})$ for any i and t. Then, for any t, the distribution of Y_{it} is governed by $\left(g(y; \beta_{i^\star}^{(1)} z_{i^\star \to i}(x_{it}^{(1)}, x_{it}^{(2)}) + \beta_{i^\star}^{(2)} x_{it}^{(2)} + \alpha_{i^\star} + \xi_t)\right)_{y=0}^{L}$. Because of Assumption 1(c)1, we then know the true value of $g(\bar{y}; (\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})z + \alpha_{i^\star} + \xi_t)$ for any $z \in \mathcal{Z}_t^{i^\star}$ and t when N and T are infinity. Assumption 3(b) ensures that we can find a t and $z, z' \in \mathcal{Z}_t^{i^\star}$ with $z \neq z'$ such that $g(\bar{y}; (\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})z + \alpha_{i^\star} + \xi_t) = g(\bar{y}; (\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})z' + \alpha_{i^\star} + \xi_t)$. Then, due to Assumption 1(b), we have $(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})z = (\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)})z'$ and identify $\beta_{i^\star}^{(2)}/\beta_{i^\star}^{(1)} = (z^{(1)} - z'^{(1)})/(z'^{(2)} - z^{(2)})$.

The second statement. Let s, t such that

$$\left\{g(\bar{y}; \beta_{i^\star}^{(1)} z + \beta_{i^\star}^{(2)} x^{(2)} + \alpha_{i^\star} + \xi_t) : \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)}, \alpha_{i^\star})}(z, x^{(2)}, 1) \in \cap_{i \in \mathbb{N} \setminus \mathcal{I}_0} \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)}, \alpha_{i^\star})}(\mathcal{Z}_{it}^{i^\star}, 1)\right\} \cap$$
$$\left\{g(\bar{y}; z + \beta_1^{(2)} x^{(2)} + \alpha_{i^\star} + \xi_s) : \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)}, \alpha_{i^\star})}(z, x^{(2)}, 1) \in \cap_{i \in \mathbb{N} \setminus \mathcal{I}_0} \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)}, \alpha_{i^\star})}(\mathcal{Z}_{is}^{i^\star}, 1)\right\} \neq \emptyset.$$

We can then find $(z, x^{(2)})$ with $\mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)}, \alpha_{i^\star})}(z, x^{(2)}, 1) \in \cap_{i \in \mathbf{N}} \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)}, \alpha_{i^\star})}(\mathcal{Z}_{it}^{i^\star}, 1)$, and $(z', x^{(2)'})$ with $\mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)}, \alpha_{i^\star})}(z', x^{(2)'}, 1) \in \cap_{i \in \mathbf{N}} \mathrm{P}_{(\beta_{i^\star}^{(1)}, \beta_{i^\star}^{(2)}, \alpha_{i^\star})}(\mathcal{Z}_{is}^{i^\star}, 1)$, such that

$$\beta_{i^\star}^{(1)} z + \beta_{i^\star}^{(2)} x^{(2)} + \alpha_{i^\star} + \xi_t = \beta_{i^\star}^{(1)} z' + \beta_{i^\star}^{(2)} x^{(2)'} + \alpha_{i^\star} + \xi_s.$$

We then identify $(\xi_s - \xi_t)/\beta_{i^\star}^{(1)}$ by $z - z' + (\beta_{i^\star}^{(2)}/\beta_{i^\star}^{(1)})(x^{(2)} - x^{(2)'})$.

The third statement. First note that we can identify $(\xi_t - \xi_{t^\star})/\beta_{i^\star}^{(1)}$ that satisfies (6) with $s = t^\star$. Following the same logic we identify $(\xi_t - \xi_{t^\star})/\beta_{i^\star}^{(1)}$ for $t \in \mathcal{T}^\star$. Since the set $\mathrm{P}_{(1,\beta_{i^\star}^{(2)}/\beta_{i^\star}^{(1)})}(\mathcal{Z}_{it}^{i^\star})$ is identified for each $i \in \mathbb{N} \setminus \mathcal{I}_0$, we then identify $g(y; \beta_{i^\star}^{(1)} u + \alpha_{i^\star} + \xi_{t^\star})$ for $(y,u) \in \mathcal{Y} \times \cap_{i\in\mathbb{N}\setminus\mathcal{I}_0} \mathrm{P}_{(1,\beta_{i^\star}^{(2)}/\beta_{i^\star}^{(1)})}(\mathcal{Z}_{it}^{i^\star}) + (\xi_t - \xi_{t^\star})/\beta_{i^\star}^{(1)}$ for any $t \in \mathcal{T}^\star$ and therefore for $(y,u) \in \mathcal{Y} \times \cup_{t\in\mathcal{T}^\star} \left(\cap_{i\in\mathbb{N}\setminus\mathcal{I}_0} \mathrm{P}_{(1,\beta_{i^\star}^{(2)}/\beta_{i^\star}^{(1)})}(\mathcal{Z}_{it}^{i^\star}) + (\xi_t - \xi_{t^\star})/\beta_{i^\star}^{(1)} \right)$. The proof is completed.

C Connectedness of $\mathcal{G}^\infty$ and Support $\mathcal{X}_{i^*}^1$

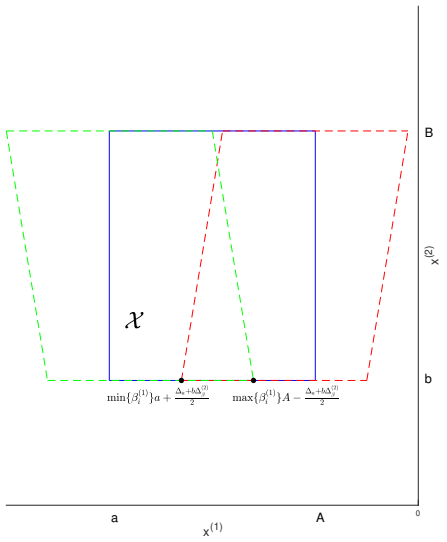
Without loss of generality, suppose $i^\star = 1$. The support of $x_1^{(1)}$, $\mathcal{X}_1^1$, plays an important role in determining the connectedness of $\mathcal{G}^\infty$ (Assumption 3(a)). When $x_1^{(1)}$ has a large support, e.g., $\mathcal{X}_1^1 = \mathbf{R}$, Assumption 3 holds trivially and $\mathcal{G}^\infty$ is connected. When $\mathcal{X}_1^1$ is not the entire real line (e.g., a box), $\mathcal{G}^\infty$ can still be connected and Assumption 3 holds. The required support condition on $\mathcal{X}_1^1$ is determined by the ranges of (and economic restrictions on) (β_i, α_i) and ξ_t. We elaborate these points in two examples.

Example 1. *Suppose that $\mathcal{X}_i = \mathcal{X} = [a,A] \times [b,B]$ where $a < A < 0$ and $0 < b < B$. Moreover, $\mathcal{Z}_{it} = \mathcal{Z}_i$ for any $(i,t) \in \mathbf{N}^2$. This setting can be considered as a demand model with $x^{(1)}$ being minus price of the goods and $x^{(2)}$ being its quality. Correspondingly, coefficient $\beta_i^{(1)} > 0$ (downward-sloping demand) is interpreted as the extent of the disutility of price and $\beta_i^{(2)}$ the preference for quality.*
In addition, suppose that $\max\{\beta_i^{(1)}\} > 1 > \min\{\beta_i^{(1)}\} > 1/\max\{\beta_i^{(1)}\} > 0$, $\frac{1}{2}\Delta_\beta^{(2)} := \max\{\beta_i^{(2)}\} - \beta_1^{(2)} = \beta_1^{(2)} - \min\{\beta_i^{(2)}\}$, and $\frac{1}{2}\Delta_\alpha := \max\{\alpha_i\} - \alpha_1 = \alpha_1 - \min\{\alpha_i\}$, i.e., individual 1's (α_1, β_1), supposedly equal to $(0,1)$ to simplify the exposition, is at the center of the range of $(\alpha_i, \beta_i) \in [\min\{\alpha_i\}, \max\{\alpha_i\}] \times [\min\{\beta_i^{(1)}\}, \max\{\beta_i^{(1)}\}] \times [\min\{\beta_i^{(2)}\}, \max\{\beta_i^{(2)}\}]$, where quantities defined by an application of the max *and* min *operators are well-defined.*
First, Assumption 3(a) holds and $\mathcal{G}^\infty$ is connected if for any (α_i, β_i), there exists $x_i \in \mathcal{X}$ such that $\alpha_i + \beta_i^{(1)} x_i^{(1)} + x_i^{(2)}(\beta_i^{(2)} - \beta_1^{(2)}) \in (a, A)$. Because of the connectedness of $\mathcal{X}$, continuity of the linear mapping $x \to z_i(x^{(1)}; x^{(2)})$, and the intermediate value theorem, this is equivalent to

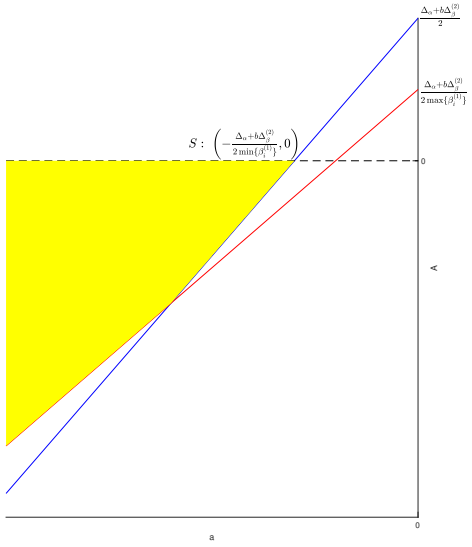
$$\begin{aligned} &\sup_{(\alpha_i,\beta_i)} \inf_{x\in\mathcal{X}} \{\alpha_i + \beta_i^{(1)} x^{(1)} + x^{(2)}(\beta_i^{(2)} - \beta_1^{(2)})\} = \max\{\alpha_i\} + \min\{\beta_i^{(1)}\} a + b(\max\{\beta_i^{(2)}\} - \beta_1^{(2)}) < A, \\ &\inf_{(\alpha_i,\beta_i)} \sup_{x\in\mathcal{X}} \{\alpha_i + \beta_i^{(1)} x^{(1)} + x^{(2)}(\beta_i^{(2)} - \beta_1^{(2)})\} = \min\{\alpha_i\} + \max\{\beta_i^{(1)}\} A + b(\min\{\beta_i^{(2)}\} - \beta_1^{(2)}) > a. \end{aligned} \tag{16}$$

Figure 3: Support Condition on $x^{(1)}$ and Connectedness of $\mathcal{G}^\infty$ in Assumption 3

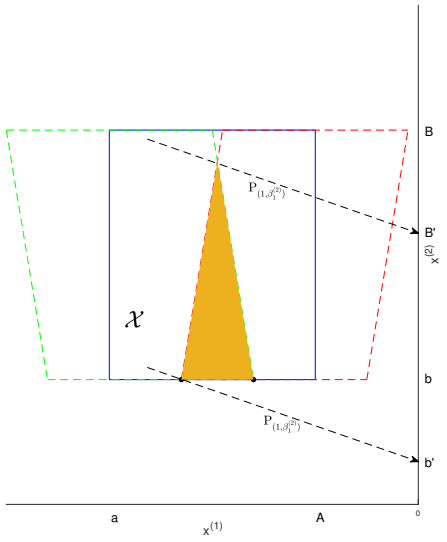
(a) Illustration of Assumption 3(a)



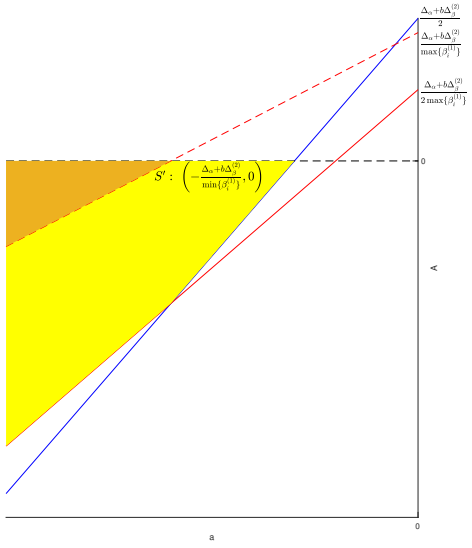
(b) (a, A) with which Assumption 3(a) holds



(c) Illustration of Assumption 3(b)



(d) (a, A) with which Assumption 3(b) holds



The geometric interpretation of (16) *is illustrated in Figure 3(a). The linear mapping* $x \to (z_i(x^{(1)}; x^{(2)}), x^{(2)})$ *maps the box* $\mathcal{X}$ *to a parallelogram that overlaps with int*$(\mathcal{X})$*, the interior of* $\mathcal{X}$ *(e.g., the red and green ones in Figure 3(a)). The first inequality in* (16) *requires the red parallelogram corresponding to the mapping defined by* $(\max\{\alpha_i\}, \min\{\beta_i^{(1)}\}, \max\{\beta_i^{(2)}\})$*, which is stretched to the right, to overlap with int*$(\mathcal{X})$*. Similarly, the second inequality requires the green paral-*

lelogram corresponding to $(\min\{\alpha_i\}, \max\{\beta_i^{(1)}\}, \min\{\beta_i^{(2)}\})$*, which is stretched to the left, to overlap with int*$(\mathcal{X})$*. These two parallelograms are the "most distant" from* $\mathcal{X}$ *in either direction. Inequalities* (16) *are further equivalent to:*

$$\begin{aligned} A &> \min\{\beta_i^{(1)}\}a + \frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{2}, \\ A &> \frac{1}{\max\{\beta_i^{(1)}\}}a + \frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{2\max\{\beta_i^{(1)}\}}. \end{aligned} \tag{17}$$

The yellow region defined by the blue line (the first inequality in (17)*) and red line (the second in* (17)*) in Figure 3(b) shows the values of* (a, A) *that satisfy* (17)*. In particular,* A *can be close to zero (the lower bound of the observed price of the goods is small) and* a *close to* $-(\Delta_\alpha + b\Delta_\beta^{(2)})/(2\min\{\beta_i^{(1)}\})$*, as illustrated by point* S*. The size of the corresponding support for* $x^{(1)}$*,* $A - a$*, is then close to* $(\Delta_\alpha + b\Delta_\beta^{(2)})/(2\min\{\beta_i^{(1)}\})$*. For any size greater than this length, Assumption 3(a) can always hold with some* (a, A)*. This minimal support requirement becomes more stringent when the ranges of* α_i (Δ_α) *and* $\beta_i^{(2)}$ $(\Delta_\beta^{(2)})$ *increase.*

Second, Assumption 3(b) holds when one further requires that the parallelograms the most distant from $\mathcal{X}$ *overlap, as illustrated by the orange region in Figure 3(c). This is because the preimage of* $\mathrm{P}_{(1,\beta_1^{(2)})}(x) = r$ *for any* $r \in (b', B')$ *is a line segment in this region and therefore not a singleton. In particular, this implies*

$$\begin{aligned} &\max\{\alpha_i\} + \min\{\beta_i^{(1)}\}a + b(\max\{\beta_i^{(2)}\} - \beta_1^{(2)}) < \min\{\alpha_i\} + \max\{\beta_i^{(1)}\}A + b(\min\{\beta_i^{(2)}\} - \beta_1^{(2)}) \\ \implies &A > \frac{\min\{\beta_i^{(1)}\}}{\max\{\beta_i^{(1)}\}}a + \frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{\max\{\beta_i^{(1)}\}}. \end{aligned} \tag{18}$$

Inequality (18) *is stronger than the second one in* (17) *and is represented by the dashed red line in Figure 3(d). The values of* (a, A) *with which Assumption 3(b) holds, the orange region in Figure 3(d), are then more limited than the yellow one (corresponding to Assumption 3(a)) and the strict lower bound of* $A - a$*,* $(\Delta_\alpha + b\Delta_\beta^{(2)})/(\min\{\beta_i^{(1)}\})$ *(achieved at* S'*), is greater than* $(\Delta_\alpha + b\Delta_\beta^{(2)})/(2\min\{\beta_i^{(1)}\})$*. In other words, identifying further* $\beta_1^{(2)}$ *requires a larger support of* $x^{(1)}$ *than the one needed for the identification of* α_i*,* $\beta_i^{(1)}$*, and* $\beta_i^{(2)} - \beta_1^{(2)}$*.*

Finally, suppose that $t^\star = 1$ *in Theorem 2. For any* t*, condition* (6) *in the second point of Theorem 2 holds when the projection of the orange region by* $\mathrm{P}_{(1,\beta_1^{(2)})}$ *(the segment between* b' *and* B' *in Figure 3(c)) intersects with itself when translated by* ξ_t *for any* $t \in \mathbf{T}$*. One can then identify* ξ_t *for* $|\xi_t| < B' - b'$*. To point identify*

ξ_t with $|\xi_t| \geq B' - b'$, one may need enlarge the support of $x^{(1)}$ relative to that required by Assumption 3(b).

The next example illustrates how shape restrictions on the dependence of parameters of interest on observed individual characteristics can attenuate support requirement on $x^{(1)}$.

Example 2 (Shape restriction and the support of $x^{(1)}$)**.** *We use the setting in Example 1 but drop $x^{(2)}$ from the model. Suppose that $\beta_i^{(1)}$, the parameter of disutility of price, is an unknown continuous function of w_i, individual i's income, denoted by $\beta(w_i)$, and decreases in w_i, i.e., a richer individual is less sensitive to price change. Moreover, α_i is a continuous function of w_i, denoted by $\alpha(w_i)$.*

Start with the individual with the highest income $\bar{w}_i$ whose $\beta(\bar{w}_i)$ is then the smallest. Now consider the individual whose income is slightly below, say $\bar{w}_i - \epsilon$. Then, this individual's $\beta(\bar{w}_i - \epsilon)$ is slightly greater than $\beta(\bar{w}_i)$ and the corresponding $\alpha(\bar{w}_i - \epsilon)$ is slightly different from $\alpha(\bar{w}_i)$. Then, the individual with the highest income can be compensated by the one with the slightly lower income if there exists $x^{(1)} \in (a, A)$,

$$\frac{\alpha(\bar{w}_i) - \alpha(\bar{w}_i - \epsilon)}{\beta(\bar{w}_i - \epsilon)} + \frac{\beta(\bar{w}_i)}{\beta(\bar{w}_i - \epsilon)} x^{(1)} \in (a, A). \tag{19}$$

Because $\frac{\alpha(\bar{w}_i) - \alpha(\bar{w}_i - \epsilon)}{\beta(\bar{w}_i - \epsilon)} \approx 0$ and $\frac{\beta(\bar{w}_i)}{\beta(\bar{w}_i - \epsilon)} \approx 1$, the compensating variable $\frac{\alpha(\bar{w}_i) - \alpha(\bar{w}_i - \epsilon)}{\beta(\bar{w}_i - \epsilon)} + \frac{\beta(\bar{w}_i)}{\beta(\bar{w}_i - \epsilon)} x^{(1)}$ is always in a neighborhood of $x^{(1)}$. Consequently, as long as $A > a$, (19) always holds.

We can repeat this argument to another individual with a slightly lower income than $\bar{w}_i - \epsilon$ and show that she is compensable by the individual with income $\bar{w}_i - \epsilon$, forming the required sequence of compensation and achieving the connectedness of $\mathcal{G}^\infty$. Note that we only require $A > a$ and the size of the support $A - a$ can be arbitrarily small.

D Proof of Theorem 3

Firstly, we prove that Theorem 2 in section 2 implies Assumption 6 in the setting of model (2) (Appendix D.1). Secondly, we propose additional conditions that are sufficient for Assumption 7 (Appendix D.2). Finally, we prove Theorem 3. In both Appendices D.1 and D.2, we suppose Assumptions 1–5 hold.

D.1 Sufficient Conditions for Assumption 6

Suppose that the conditions in Theorem 2 hold. Suppose now there is another maximizer $(G_1', G_2', (h_l')_{l=1}^L) \in \mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta})) \times \mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi)) \times \bar{\Theta}_h$. Then, the following equality holds almost surely in terms of the joint distribution of $(\alpha', \beta', \xi', x, \alpha, \beta, \xi) \sim G_1'(\alpha', \beta'; \alpha, \beta) \times G_2'(\xi'; \xi) \times F_0(x, \alpha, \beta, \xi)$: for $y = 1, \ldots, L$,

$$g'(y; \alpha' + \xi' + x'\beta') = g_0(y; \alpha + \xi + x'\beta). \tag{20}$$

Without loss of generality, suppose $g'(1; v)$ and $g_0(1; v)$ satisfy the strict monotonicity condition in Assumption 1(b). Then, the following equality holds almost surely:

$$\alpha' + \xi' + x'\beta' = (g')^{-1}(1; g_0(1; \alpha + \xi + x'\beta)).$$

Now fixing $(x, \alpha', \beta', \alpha, \beta)$ to some values $(\tilde{x}, \tilde{\alpha}', \tilde{\beta}', \tilde{\alpha}, \tilde{\beta})$ in the domain, we obtain:

$$\xi' = (g')^{-1}(1; g_0(1; \tilde{\alpha} + \xi + \tilde{x}'\tilde{\beta})) - \tilde{\alpha}' - \tilde{x}'\tilde{\beta}'$$

almost surely for (ξ', ξ). Consequently, ξ' is a function of ξ. Denote this function by $\phi_3(\xi)$.

Now using Assumption 3(i), we can find $(x^{(1)m}, x^{(2)m})$, $m = 1, 2, 3$, in the domain of X such that

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix}$$

is of full rank. Then, fixing ξ to some value $\tilde{\xi}$ in the domain, we have:

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix} \begin{bmatrix} \alpha' \\ \beta' \end{bmatrix} = \begin{bmatrix} (g')^{-1}(1; g_0(1; \alpha + \tilde{\xi} + (x^{(1)1}, x^{(2)1})\beta)) - \phi_3(\tilde{\xi}) \\ (g')^{-1}(1; g_0(1; \alpha + \tilde{\xi} + (x^{(1)2}, x^{(2)2})\beta)) - \phi_3(\tilde{\xi}) \\ (g')^{-1}(1; g_0(1; \alpha + \tilde{\xi} + (x^{(1)3}, x^{(2)3})\beta)) - \phi_3(\tilde{\xi}) \end{bmatrix}.$$

We then obtain that (α', β') is a vector of functions of (α, β), denoted by $\phi_1(\alpha, \beta)$ and $\phi_2(\alpha, \beta)$ respectively. We plug ϕ_3, ϕ_1, and ϕ_2 to (20) and obtain: for any $y = 1, \ldots, L$, almost surely

$$g'(y; \phi_1(\alpha, \beta) + \phi_3(\xi) + x'\phi_2(\alpha, \beta)) = g_0(y; \alpha + \xi + x'\beta).$$

In other words, a model (2) with $(\alpha_i, \beta_i) = (\phi_1(\alpha_{i0}, \beta_{i0}), \phi_2(\alpha_{i0}, \beta_{i0}))$ and

$\xi_t = \phi_3(\xi_{t0})$ for all i and t, and $g(y;\cdot) = g'(y;\cdot)$ for all $y = 1,\ldots,L$ is equivalent to the true model. Note that the normalizations also apply to $(\phi_1(\alpha_{i^\star 0},\beta_{i^\star 0}),(1,0)\phi_2(\alpha_{i^\star 0},\beta_{i^\star 0}),\phi_3(\xi_{t^\star 0}))$. Then, because of Theorem 2, we obtain that $\phi_1(\alpha_{i0},\beta_{i0}) = \alpha_{i0}$, $\phi_2(\alpha_{i0},\beta_{i0}) = \beta_{i0}$, and $\phi_3(\xi_{t0}) = \xi_{t0}$. Consequently, $g'(y;\cdot) = g(y;\cdot)$ for all $y = 1,\ldots,L$. The uniqueness is proved.

D.2 Sufficient Conditions for Assumption 7

Define

$$A_h(x) = \sup_{(\alpha,\beta,\xi)\in\Omega_{(\alpha,\beta)}\times\Omega_\xi, 0\le l\le L} \left\{ \left| \log \frac{\exp\{h_l(\alpha+\xi+x^{\mathrm{T}}\beta)\}}{1+\sum_{l=1}^L \exp\{h_l(\alpha+\xi+x^{\mathrm{T}}\beta)\}} \right| \right\}.$$

Condition 1.

(a) $\frac{\sum_{i=1}^N \sum_{t=1}^T \|x_{it}\|}{NT} = O_p(1)$.

(b) $\sup_{h\in\bar\Theta_h,(\alpha,\beta,\xi)\in\Omega_{(\alpha,\beta)}\times\Omega_\xi} \mathbb{E}\left[A_h^2(X)|\alpha,\beta,\xi\right] \le M_1 < \infty$.

(c) $\sup_{(\alpha,\beta,\xi,x,h)\in\Omega_{(\alpha,\beta)}\times\Omega_\xi\times\Omega_x\times\bar\Theta_h} \mathbb{E}\left[h_Y^4(\alpha+\xi+x^T\beta)\right] \le M_2$ *for some* $1 < M_2 < \infty$.

(d) $\sup_{1\le y\le L, h_y\in\bar\Theta_{hy},v} |\partial_v h_y(v)| \le M_3$.

Condition 1(a)–(d) impose regularities on the distribution of X_{it} and the log-likelihood function. When X_{it} has bounded support and the norm $\|\cdot\|_h$ is stronger than the sup norm on h and its derivatives, Condition 1(a) holds and Conditions 1(b)–1(d) can be implied by Assumptions 1, 4, and 5.

Assumption 7(a). For any $\eta > 0$, consider a collection of open balls $\{\mathcal{B}(\alpha,\beta;\eta/(8M_3M_x)) : (\alpha,\beta)\in\Omega_{\alpha,\beta}\}$ where $M_x \ge 1$ satisfies $\frac{\sum_{i,t}\|(1,x_{it}')\|}{NT} \le M_x$ asymptotically wp1. Because $\Omega_{\alpha,\beta}$ is compact, we can then find $\mathcal{S}_m^{\alpha,\beta} := \mathcal{B}(\alpha^{(m)},\beta^{(m)};\eta/(8M_3M_x)), m = 1,\ldots,M_\eta$ such that $\Omega_{\alpha,\beta}\subset\cup_{m=1}^{M_\eta}\mathcal{S}_m^{\alpha,\beta}$. Similarly, we can find a finite number of open balls $\mathcal{S}_m^\xi := \mathcal{B}(\xi^{(n)};\eta/(8M_3)), n = 1,\ldots,N_\eta$ such that $\Omega_\xi \subset \cup_{n=1}^{N_\eta}\mathcal{S}_m^\xi$ and $\mathcal{S}_r^h := \mathcal{B}(h^{(r)};\eta/8), r = 1,\ldots,R_\eta$ such that $\bar\Theta_h \subset \cup_{r=1}^{R_\eta}\mathcal{S}_m^h$. Define

$$(\alpha,\beta)\in^*\mathcal{S}_m^{\alpha,\beta} \iff m = \text{the minimal } m' \text{ such that } (\alpha,\beta)\in\mathcal{S}_{m'}^{\alpha,\beta}.$$

Then, for any $(\alpha, \beta) \in \Omega_{\alpha,\beta}$, there exists a unique m such that $(\alpha_i, \beta_i) \in^* \mathcal{S}_m^{\alpha,\beta}$. Similarly, we can define such a relationship $\in^*$ for $\xi \in \Omega_\xi$ and $h \in \bar{\Theta}_h$.
Define

$$\begin{aligned}\Delta h_{it}^{(r)}(m,n) &= \log\left(\frac{\exp\{h_{y_{it}}^{(r)}(\alpha^{(m)} + \xi^{(n)} + x_{it}'\beta^{(m)})\}}{1 + \sum_{y=1}^{L}\exp\{h_y^{(r)}(\alpha^{(m)} + \xi^{(n)} + x_{it}'\beta^{(m)})\}}\right) \\ &\quad - \mathbb{E}_0\left[\log\left(\frac{\exp\{h_{y_{it}}^{(r)}(\alpha^{(m)} + \xi^{(n)} + x_{it}'\beta^{(m)})\}}{1 + \sum_{y=1}^{L}\exp\{h_y^{(r)}(\alpha^{(m)} + \xi^{(n)} + x_{it}'\beta^{(m)})\}}\right)\middle| x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0}\right] \\ &= h_{y_{it}}^{(r)}(\alpha^{(m)} + \xi^{(n)} + x_{it}'\beta^{(m)}) - \mathbb{E}_0\left[h_{y_{it}}^{(r)}(\alpha^{(m)} + \xi^{(n)} + x_{it}'\beta^{(m)})|x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0}\right],\end{aligned}$$

and

$$\Delta h_{it} = h_{y_{it}}(\alpha_i + \xi_t + x_{it}'\beta_i) - \mathbb{E}_0\left[h_{y_{it}}(\alpha_i + \xi_t + x_{it}'\beta_i)|x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0}\right].$$

Then,

$$\begin{aligned}&\mathscr{L}_{NT}(\theta_1^{NT}, h) - \mathscr{L}_{NT}^0(\theta_1^{NT}, h) \\ &\quad = \frac{1}{NT}\sum_{i,t}\sum_{r=1}^{R_\eta}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\mathbf{1}\{(\alpha_i, \beta_i) \in^* \mathcal{S}_m^{\alpha,\beta}\}\mathbf{1}\{\xi_t \in^* \mathcal{S}_n^{\xi}\}\mathbf{1}\{h \in^* \mathcal{S}_r^{h}\}\Delta h_{it}^{(r)}(m,n) \\ &\qquad + \frac{1}{NT}\sum_{i,t}\sum_{r=1}^{R_\eta}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\mathbf{1}\{(\alpha_i, \beta_i) \in^* \mathcal{S}_m^{\alpha,\beta}\}\mathbf{1}\{\xi_t \in^* \mathcal{S}_n^{\xi}\}\mathbf{1}\{h \in^* \mathcal{S}_r^{h}\}\left(\Delta h_{it} - \Delta h_{it}^{(r)}(m,n)\right).\end{aligned} \tag{21}$$

Moreover, for any $(\alpha, \beta, \xi, h, x, y)$ and $(\tilde{\alpha}, \tilde{\beta}, \tilde{\xi}, \tilde{h}, x, y)$, we have

$$\begin{aligned}&\left|h_y(\alpha + \xi + x'\beta) - \tilde{h}_y(\tilde{\alpha} + \tilde{\xi} + x'\tilde{\beta})\right| \\ &\quad \le \left|h_y(\alpha + \xi + x'\beta) - h_y(\tilde{\alpha} + \tilde{\xi} + x'\tilde{\beta})\right| + \left|h_y(\tilde{\alpha} + \tilde{\xi} + x'\tilde{\beta}) - \tilde{h}_y(\tilde{\alpha} + \tilde{\xi} + x'\tilde{\beta})\right| \\ &\quad \le \sup_v|\partial h_y(v)| \times \left|\alpha - \tilde{\alpha} + \xi - \tilde{\xi} + x'(\beta - \tilde{\beta})\right| + \|h - \tilde{h}\|_h \\ &\quad \le M_3(\|(1, x')\| \times \|(\alpha - \tilde{\alpha}, \beta - \tilde{\beta})\| + |\xi - \tilde{\xi}|) + \|h - \tilde{h}\|_h.\end{aligned}$$

Similarly,

$$\begin{aligned}&\left|\mathbb{E}_0\left[h_y(\alpha_i + \xi_t + x'\beta_i) - \tilde{h}_y(\tilde{\alpha} + \tilde{\xi} + x'\tilde{\beta})|x, \alpha_{i0}, \beta_{i0}, \xi_{t0}\right]\right| \\ &\quad \le M_3(\|(1, x')\| \times \|(\alpha_i - \tilde{\alpha}, \beta_i - \tilde{\beta})\| + |\xi_t - \tilde{\xi}|) + \|h - \tilde{h}\|_h.\end{aligned}$$

Then, the absolute value of the third line in (21) can be bounded by

$\frac{2}{NT}\sum_{i,t}\left[M_3\left(\|(1,x'_{it})\|\eta/(8M_3M_x)+\eta/(8M_3)\right)+\eta/8\right]\leq\frac{3\eta}{4}$. Then,

$$\begin{aligned}&\Pr\left(\sup_{\theta_1^{NT}\in\Omega_{\alpha,\beta}^N\times\Omega_\xi^T,h\in\bar{\Theta}_h}\left|\mathscr{L}_{NT}(\theta_1^{NT},h)-\mathscr{L}_{NT}^0(\theta_1^{NT},h)\right|>\eta\right)\\&\leq\Pr\left(\sup_{(\alpha_i,\beta_i)\in\Omega_{\alpha,\beta},\xi_t\in\Omega_\xi,1\leq i\leq N,1\leq t\leq T,h\in\bar{\Theta}_h}\left|\sum_{r=1}^{R_\eta}\mathbf{1}\{h\in^*\mathcal{S}_r^h\}\frac{1}{NT}\sum_{i,t}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\mathbf{1}\{(\alpha_i,\beta_i)\in^*\mathcal{S}_m^{\alpha,\beta}\}\mathbf{1}\{\xi_t\in^*\mathcal{S}_n^\xi\}\Delta h_{it}^{(r)}(m,n)\right|>\frac{\eta}{4}\right).\end{aligned}$$

We now show that for each $r=1,\ldots,R_\eta$,

$$\Pr\left(\sup_{(\alpha_i,\beta_i)\in\Omega_{\alpha,\beta},\xi_t\in\Omega_\xi,1\leq i\leq N,1\leq t\leq T}\left|\frac{1}{NT}\sum_{i,t}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\mathbf{1}\{(\alpha_i,\beta_i)\in^*\mathcal{S}_m^{\alpha,\beta}\}\mathbf{1}\{\xi_t\in^*\mathcal{S}_n^\xi\}\Delta h_{it}^{(r)}(m,n)\right|>\frac{\eta}{4R_\eta}\right)\to 0. \tag{22}$$

This will imply $\Pr\left(\sup_{\theta_1^{NT}\in\Omega_{\alpha,\beta}^N\times\Omega_\xi^T,h\in\bar{\Theta}_h}\left|\mathscr{L}_{NT}(\theta_1^{NT},h)-\mathscr{L}_{NT}^0(\theta_1^{NT},h)\right|>\eta\right)\to 0$ and therefore Assumption 7(a).

Define

$$\lambda_{NT}^{(r)}(m,n)=\sup_{u\in\mathbb{R}^N,v\in\mathbb{R}^T,\|u\|=\|v\|=1}|u^{\mathrm{T}}(\Delta h_{it}^{(r)})_{i,t}(m,n)v|.$$

In other words, $\lambda_{NT}^{(r)}(m,n)$ is the greatest absolute value of the singular values of the matrix $(\Delta h_{it}^{(r)})_{i,t}(m,n)$ given N and T. Using Conditions 1(c) and Theorem 2 of Latała (2005), we have $\mathbb{E}_0\left[\lambda_{NT}^{(r)}(m,n)\right]\leq C\sqrt{M_2}\max\{\sqrt{N},\sqrt{T}\}$, where C is a constant that does not depend on N, T, n, mr. Then, for any $(\alpha_i,\beta_i)\in\Omega_{\alpha,\beta},\xi_t\in\Omega_\xi,1\leq i\leq N,1\leq t\leq T$,

$$\begin{aligned}&\left|\frac{1}{NT}\sum_{i,t}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\mathbf{1}\{(\alpha_i,\beta_i)\in^*\mathcal{S}_m^{\alpha,\beta}\}\mathbf{1}\{\xi_t\in^*\mathcal{S}_n^\xi\}\Delta h_{it}^{(r)}(m,n)\right|\\&=\left|\frac{1}{NT}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}(\mathbf{1}\{(\alpha_i,\beta_i)\in^*\mathcal{S}_m^{\alpha,\beta}\})_{i=1}^N(\Delta h_{it}^{(r)}(m,n))_{i,t}[(\mathbf{1}\{\xi_t\in^*\mathcal{S}_n^\xi\})_{t=1}^T]^{\mathrm{T}}\right|\\&\leq\frac{1}{NT}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\lambda_{NT}^{(r)}(m,n)\|(\mathbf{1}\{(\alpha_i,\beta_i)\in^*\mathcal{S}_m^{\alpha,\beta}\})_{i=1}^N\|\times\|(\mathbf{1}\{\xi_t\in\mathcal{S}_n^\xi\})_{t=1}^T\|\\&\leq\frac{1}{\sqrt{NT}}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\lambda_{NT}^{(r)}(m,n).\end{aligned}$$

Then,

$$
\begin{aligned}
&\Pr\left(\sup_{(\alpha_i,\beta_i)\in\Omega_{\alpha,\beta},\xi_t\in\Omega_\xi,1\le i\le N,1\le t\le T}\left|\frac{1}{NT}\sum_{i,t}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\mathbf{1}\{(\alpha_i,\beta_i)\in^* \mathcal{S}_m^{\alpha,\beta}\}\mathbf{1}\{\xi_t\in^*\mathcal{S}_n^{\xi}\}\Delta h_{it}^{(r)}(m,n)\right|>\frac{\eta}{4R_\eta}\right)\\
\le&\Pr\left(\frac{1}{\sqrt{NT}}\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\lambda_{NT}^{(r)}(m,n)>\frac{\eta}{4R_\eta}\right)\\
\le&\frac{4R_\eta\mathbb{E}_0\left[\sum_{m=1}^{M_\eta}\sum_{n=1}^{N_\eta}\lambda_{NT}^{(r)}(m,n)\right]}{\eta\sqrt{NT}}\\
\le&\frac{4R_\eta M_\eta N_\eta C\sqrt{M_2}\max\{\sqrt{N},\sqrt{T}\}}{\eta\sqrt{NT}}\to 0.
\end{aligned}
$$

This verifies (22) and Assumption 7(a).

Assumption 7(c). We verify Assumptions 2 and 3A in Newey (1991) and then use Corollary 2.2 in the same paper to prove Assumption 7(c).

For (G',h') and (G,h), we have

$$
\begin{aligned}
&\left|\mathscr{L}^0(G,h,F_0)-\mathscr{L}^0(G',h',F_0)\right|\\
&\le\sum_{y=0}^{L}\int\frac{\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}{1+\sum_{y=1}^{L}\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}\\
&\quad\times\left[\int\left|\log\left(\frac{\exp\{h_y(\alpha'+\xi'+x'\beta')\}}{1+\sum_{y=1}^{L}\exp\{h_y(\alpha'+\xi'+x'\beta')\}}\right)\right|(dG_1(\alpha',\beta'|\alpha,\beta)-dG_1'(\alpha',\beta'|\alpha,\beta))dG_2(\xi'|\xi)\right]\\
&\quad\times dF_0(x,\alpha,\beta,\xi)\\
&\quad+\sum_{y=0}^{L}\int\frac{\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}{1+\sum_{y=1}^{L}\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}\\
&\quad\times\left[\int\left|\log\left(\frac{\exp\{h_y(\alpha'+\xi'+x'\beta')\}}{1+\sum_{y=1}^{L}\exp\{h_y(\alpha'+\xi'+x'\beta')\}}\right)\right|(dG_2(\xi'|\xi)-dG_2'(\xi'|\xi))dG_1'(\alpha',\beta'|\alpha,\beta)\right]\\
&\quad\times dF_0(x,\alpha,\beta,\xi)\\
&\quad+\sum_{y=0}^{L}\int\frac{\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}{1+\sum_{y=1}^{L}\exp\{h_{y0}(\alpha+\xi+x'\beta)\}}\\
&\quad\times\left[\int\left|\log\left(\frac{\exp\{h_y(\alpha'+\xi'+x'\beta')\}}{1+\sum_{y=1}^{L}\exp\{h_y(\alpha'+\xi'+x'\beta')\}}\right)-\log\left(\frac{\exp\{h_y'(\alpha'+\xi'+x'\beta')\}}{1+\sum_{y=1}^{L}\exp\{h_y'(\alpha'+\xi'+x'\beta')\}}\right)\right|\right.\\
&\quad\left.\times dG_2'(\xi'|\xi)dG_1'(\alpha',\beta'|\alpha,\beta)\right]dF_0(x,\alpha,\beta,\xi)\\
&\le(|\Omega_{\alpha,\beta}|\times\|G_1-G_1'\|_1+|\Omega_\xi|\times\|G_2-G_2'\|_2)\,\mathbb{E}\left[A_h(X)\right]+2\|h-h'\|_h\\
&\le\max\{|\Omega_{\alpha,\beta}|\sqrt{M_1},|\Omega_\xi|\sqrt{M_1},2\}\max\{\|G_1-G_1'\|_1,\|G_2-G_2'\|_2,\|h-h'\|_h\}.
\end{aligned}
\tag{23}
$$

Similarly, we replace F_0 in (23) by F_0^{NT} and obtain:

$$\begin{aligned}
&\left|\mathscr{L}^0(G,h,F_0^{NT}) - \mathscr{L}^0(G',h',F_0^{NT})\right| \\
&\quad \leq \max\left\{|\Omega_{\alpha,\beta}|\frac{\sum_{i,t} A_h(x_{it})}{NT}, |\Omega_\xi|\frac{\sum_{i,t} A_h(x_{it})}{NT}, 2\right\} \max\{\|G_1 - G_1'\|_1, \|G_2 - G_2'\|_2, \|h - h'\|_h\}.
\end{aligned}$$

Under Assumption 1(a), Conditions 1(a), and 1(e), we have $\sum_{i,t} A_h(x_{it})/NT = O_p(1)$ and therefore $\max\left\{|\Omega_{\alpha,\beta}|\frac{\sum_{i,t} A_h(x_{it})}{NT}, |\Omega_\xi|\frac{\sum_{i,t} A_h(x_{it})}{NT}, 2\right\} = O_p(1)$. This implies Assumption 3A in Newey (1991).

For any $(G,h) \in \Theta_1 \times \Theta_2 \times \bar{\Theta}_h$, we have

$$\begin{aligned}
&\mathscr{L}^0(G,(h_l)_{l=1}^L, F_0^{NT}) \\
&\quad = \frac{1}{NT}\sum_{i,t}\sum_{y=0}^{L} \frac{\exp\{h_{y0}(\alpha_{i0} + \xi_{t0} + x_{it}'\beta_{i0})\}}{1 + \sum_{l=1}^L \exp\{h_{l0}(\alpha_{i0} + \xi_{t0} + x_{it}'\beta_{i0})\}} \\
&\qquad \times \int \left[\log\left(\frac{\exp\{h_y(\alpha' + \xi' + x_{it}'\beta')\}}{1 + \sum_{l=1}^L \exp\{h_l(\alpha' + \xi' + x'\beta')\}}\right)\right] dG_1(\alpha',\beta'|\alpha_{i0},\beta_{i0}) dG_2(\xi'|\xi_{t0}) \\
&\quad =: \sum_{i,t} \tau_{it},
\end{aligned}$$

with

$$\begin{aligned}
&\tau_{it} := \tau(x_{it},\alpha_{i0},\beta_{i0},\xi_{t0}) \\
&\quad \leq \sum_{y=0}^{L} \frac{\exp\{h_{y0}(\alpha_{i0} + \xi_{t0} + x_{it}'\beta_{i0})\}}{1 + \sum_{l=1}^L \exp\{h_{l0}(\alpha_{i0} + \xi_{t0} + x_{it}'\beta_{i0})\}} \log \frac{\exp\{h_{y0}(\alpha_{i0} + \xi_{t0} + x_{it}'\beta_{i0})\}}{1 + \sum_{l=1}^L \exp\{h_{l0}(\alpha_{i0} + \xi_{t0} + x_{it}'\beta_{i0})\}} \\
&\quad < 0.
\end{aligned}$$

We now evaluate $\Pr\left(\left|\frac{1}{NT}\sum_{i,t}\left(\tau(x_{it},\alpha_{i0},\beta_{i0},\xi_{t0}) - \mathbb{E}[\tau(x_{it},\alpha_{i0},\beta_{i0},\xi_{t0})]\right)\right| > \varepsilon\right)$. This amounts to evaluating the covariance between τ_{it} and $\tau_{i't'}$ for any (i,t) and (i',t'). Note that under Assumption 1,

$$\begin{aligned}
\text{for } i \neq i', t \neq t' : \ \mathrm{Cov}(\tau_{it},\tau_{i't'}) &= \mathbb{E}\left[\mathrm{Cov}(\tau_{it},\tau_{i't'}|\mathcal{F})\right] + \mathrm{Cov}(\mathbb{E}\left[\tau_{it}|\mathcal{F}\right], \mathbb{E}\left[\tau_{i't'}|\mathcal{F}\right]) \\
&= \mathrm{Cov}(\mathbb{E}\left[\tau_{it}|\mathcal{F}\right], \mathbb{E}\left[\tau_{i't'}|\mathcal{F}\right]) \\
&= \mathrm{Cov}(\mathbb{E}\left[\tau_{it}|\alpha_{i0},\beta_{i0},\xi_{t0}\right], \mathbb{E}\left[\tau_{i't'}|\alpha_{i'0},\beta_{i'0},\xi_{t'0}\right]), \\
\text{for } i = i' \text{or } t = t' : \ \mathrm{Cov}(\tau_{it},\tau_{i't'}) &= \mathbb{E}\left[\tau_{it}\tau_{i't'}\right] - \mathbb{E}\left[\tau_{it}\right]\mathbb{E}\left[\tau_{i't'}\right] \leq \mathbb{E}\left[A_h(X_{it})A_h(X_{it'})\right].
\end{aligned}$$

Note that $(\mathbb{E}\left[\tau_{it}|\alpha,\beta,\xi_{t0}\right])_{\alpha,\beta}$ define a set of measurable functions of ξ_{t0} indexed by $(\alpha,\beta) \in \Omega_{\alpha,\beta}$. Then, using Rio (1993) and Assumption 4, we have: given any

(i, i'),

$$
\begin{aligned}
&|\mathrm{Cov}(\mathbb{E}\left[\tau_{it}|\alpha_{i0}, \beta_{i0}, \xi_{t0}\right], \mathbb{E}\left[\tau_{i't'}|\alpha_{i'0}, \beta_{i'0}, \xi_{t'0}\right])| \\
&\quad \leq 2\int_0^{2a(|t-t'|)} F^{-1}_{|\mathbb{E}[\tau_{it}|\alpha_{i0},\beta_{i0},\xi_{t0}]|}(u) F^{-1}_{|\mathbb{E}[\tau_{i't'}|\alpha_{i'0},\beta_{i'0},\xi_{t'0}]|}(u) du,
\end{aligned}
$$

where $F_Z^{-1}(\cdot)$ refers to the quantile function of the random variable Z. Note that $|\mathbb{E}\left[\tau_{it}|\alpha_{i0}, \beta_{i0}, \xi_{t0}\right]| \leq \mathbb{E}\left[A_h(X_{it})|\alpha_{i0}, \beta_{i0}, \xi_{t0}\right]$. Moreover, the latter is smaller or equal to $\sqrt{M_1}$ according to Condition 1(b). Then,

$$
\begin{aligned}
&|\mathrm{Cov}(\mathbb{E}\left[\tau_{it}|\alpha_{i0}, \beta_{i0}, \xi_{t0}\right], \mathbb{E}\left[\tau_{i't'}|\alpha_{i'0}, \beta_{i'0}, \xi_{t'0}\right])| \\
\leq& 2\int_0^{2a(|t-t'|)} F^{-1}_{\mathbb{E}[A_h(X_{it})|\alpha_{i0},\beta_{i0},\xi_{t0}]}(u) F^{-1}_{\mathbb{E}[A_h(X_{i't'})|\alpha_{i'0},\beta_{i'0},\xi_{t'0}]}(u) du \\
\leq& 4a(|t-t'|)M_1.
\end{aligned}
$$

Then, since $a(m) = O(m^{-\mu})$ and therefore $a(m) \leq Cm^{-\mu}$ for some $C > 0$, we have:

$$
\begin{aligned}
&\Pr\left(\left|\frac{1}{NT}\sum_{i,t}\left(\tau(x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0}) - \mathbb{E}[\tau(x_{it}, \alpha_{i0}, \beta_{i0}, \xi_{t0})]\right)\right| > \varepsilon\right) \\
&\leq \frac{1}{\varepsilon^2 N^2 T^2}\left[\sum_{i=i' \text{ or } t=t'} M_1 + \sum_{t\neq t'} a(|t-t'|)\sum_{i\neq i'} M_1\right] \\
&\leq \frac{1}{\varepsilon^2 N^2 T^2}\left[(N^2T + T^2N)M_1 + C\sum_{t\neq t'}|t-t'|^{-\mu}(N^2-N)M_1\right] \\
&\leq \frac{M_1}{\varepsilon^2 N^2 T^2}\left[\left(1 + C\sum_{r=1}^{\infty} r^{-\mu}\right)N^2T + T^2N\right] \\
&\overset{N,T\to\infty}{\to} 0.
\end{aligned}
$$

Then, $\mathscr{L}^0(G, h, F_0^{NT}) \xrightarrow{p} \mathscr{L}^0(G, h, F_0)$ for any (G, h). Assumption 2 of Newey (1991) is verified.

Under Condition 1(b), we have $\mathbb{E}\left[\sum_{i,t} A_h(x_{it})/NT\right] \leq \sqrt{M_1}$. Then, $\mathbb{E}\left[\max\left\{|\Omega_{\alpha,\beta}|\frac{\sum_{i,t}A_h(x_{it})}{NT}, |\Omega_\xi|\frac{\sum_{i,t}A_h(x_{it})}{NT}, 2\right\}\right]$ is bounded. Using Corollary 2.2 (and the discussion below this corollary) in Newey (1991), we obtain Assumption 7(c).

Assumption 7(b). First, define

$$
\tilde{\xi}_t^*\left(\{\alpha_i, \beta_i\}_{i=1}^N, h\right) := \arg\max_{\xi\in\Omega_\xi} \underbrace{\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\cdot|\xi_{t0}}\left[\log\left(\frac{\exp\{h_{Y_{it}}(\alpha_i + \xi + X_{it}'\beta_i)\}}{1 + \sum_{y=1}^L \exp\{h_y(\alpha_i + \xi + X_{it}'\beta_i)\}}\right)\right]}_{=:\mathcal{L}_t(\xi;\{\alpha_i,\beta_i\}_{i=1}^N,h)}, \tag{24}
$$

where $\mathbb{E}_{\cdot|\xi_{t0}}[\cdot]$ is with respect to $(Y_{it}, X_{it}, \alpha_{i0}, \beta_{i0})$ conditional on ξ_{t0}. Both $\mathcal{L}_t(\xi; \{\alpha_i, \beta_i\}_{i=1}^N, h)$ and $\tilde{\xi}_t^*\left(\{\alpha_i, \beta_i\}_{i=1}^N, h\right)$ depend on t via ξ_{t0}. Similarly, define

$$(\tilde{\alpha}_i^*, \tilde{\beta}_i^*)\left(\{\xi_t\}_{t=1}^T, h\right) := \underset{(\alpha,\beta)\in\Omega_{(\alpha,\beta)}}{\arg\max} \underbrace{\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\cdot|\alpha_{i0},\beta_{i0}}\left[\log\left(\frac{\exp\{h_{Y_{it}}(\alpha+\xi_t+X_{it}'\beta)\}}{1+\sum_{y=1}^L \exp\{h_y(\alpha+\xi_t+X_{it}'\beta)\}}\right)\right]}_{=:\mathcal{L}_i(\alpha,\beta;\{\xi_t\}_{t=1}^T,h)},$$

where $\mathbb{E}_{\cdot|\alpha_{i0},\beta_{i0}}[\cdot]$ is with respect to $(Y_{it}, X_{it}, \xi_{t0})$ conditional on $(\alpha_{i0}, \beta_{i0})$.

Condition 2.

(a) For any t, $(\alpha, \beta) \in \Omega_{(\alpha,\beta)}$ and $h \in \bar{\Theta}_h$, $\mathbb{E}_{\cdot|\xi_{t0}}\left[\log\left(\frac{\exp\{h_{Y_{it}}(\alpha+\xi+X_{it}'\beta)\}}{1+\sum_{y=1}^L \exp\{h_y(\alpha+\xi+X_{it}'\beta)\}}\right)\right]$ is strongly concave with respect to ξ, uniformly with some constant $\lambda < 0$.

(b) For any i, $\xi \in \Omega_\xi$ and $h \in \bar{\Theta}_h$, $\mathbb{E}_{\cdot|\alpha_{i0},\beta_{i0}}\left[\log\left(\frac{\exp\{h_{Y_{it}}(\alpha+\xi+X_{it}'\beta)\}}{1+\sum_{y=1}^L \exp\{h_y(\alpha+\xi+X_{it}'\beta)\}}\right)\right]$ is strongly concave with respect to (α, β), uniformly with some constant $\lambda < 0$.

(c) $\partial_{(\alpha,\beta,\xi)} f_{0x}(x|\alpha, \beta, \xi)$ is bounded by some constant C_f uniformly for (x, α, β, ξ), where $f_{0x}(x|\alpha, \beta, \xi)$ is the density function of x conditional on (α, β, ξ).

We verify the stochastic equicontinuity for $\tilde{\xi}$ using Conditions 2(a) and (c). One can apply similar arguments to show the the stochastic equicontinuity of $(\tilde{\alpha}, \tilde{\beta})$ using Conditions 2(b) and (c).

Because of the strong concavity in Condition 2(b), $\tilde{\xi}_t^*\left(\{\alpha_i, \beta_i\}_{i=1}^N, h\right)$ is the unique maximizer. Suppose that $\tilde{\xi}_t^*\left(\{\alpha_i, \beta_i\}_{i=1}^N, h\right)$ is an interior solution. Using arguments similar to those in the proof of Assumption 7(a), we can show that

$$\sup_{(\alpha_i,\beta_i)\in\Omega_{\alpha,\beta},1\le i\le N;\xi\in\Omega_\xi;h\in\bar{\Theta}_h;1\le t\le T}\Bigg|\frac{1}{N}\sum_{i=1}^N \log\left[\frac{\exp\{h_{y_{it}}(\alpha_i+\xi+x_{it}'\beta_i)\}}{1+\sum_{y=1}^L \exp\{h_y(\alpha_i+\xi+x_{it}'\beta_i)\}}\right]$$
$$-\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\log\left(\frac{\exp\{h_{Y_{it}}(\alpha_i+\xi+X_{it}'\beta_i)\}}{1+\sum_{y=1}^L \exp\{h_y(\alpha_i+\xi+X_{it}'\beta_i)\}}\right)\right]\Bigg| \xrightarrow{p} 0.$$

Then,

$$\sup_{(\alpha_i,\beta_i)\in\Omega_{\alpha,\beta},1\le i\le N;h\in\bar{\Theta}_h;1\le t\le T}\left|\tilde{\xi}\left((x_{it}, y_{it})_{i=1}^N; \{\alpha_i, \beta_i\}_{i=1}^N, h\right) - \tilde{\xi}_t^*\left(\{\alpha_i, \beta_i\}_{i=1}^N, h\right)\right| \xrightarrow{p} 0.$$

Using the First-Order Condition corresponding to (24), we obtain:

$$\frac{\partial\tilde{\xi}_t^*\left(\{\alpha_i, \beta_i\}_{i=1}^N, h\right)}{\partial\xi_{t0}} = -\left[\frac{\partial^2\mathcal{L}_t(\xi; \{\alpha_i, \beta_i\}_{i=1}^N, h)}{\partial\xi^2}\right]^{-1}\frac{\partial^2\mathcal{L}_t(\xi; \{\alpha_i, \beta_i\}_{i=1}^N, h)}{\partial\xi_{t0}\partial\xi}. \quad (25)$$

Condition 2(b) implies $\left|\left[\frac{\partial^2 \mathcal{L}_t(\xi;\{\alpha_i,\beta_i\}_{i=1}^N,h)}{\partial \xi^2}\right]^{-1}\right| \leq \lambda^{-1}$. Moreover, because of Assumption 4, we have: for any (α,β,h),

$$
\begin{aligned}
&\partial^2_{\xi_{t0}\xi}\mathbb{E}_{\cdot|\xi_{t0}}\left[\log\left(\frac{\exp\{h_Y(\alpha+\xi+X'\beta)\}}{1+\sum_{y=1}^L \exp\{h_y(\alpha+\xi+X'\beta)\}}\right)\right] \\
=&\partial_{\xi_{t0}}\mathbb{E}_{\cdot|\xi_{t0}}\underbrace{\left[\partial h_Y(\alpha+\xi+X'\beta)-\sum_{l=1}^L \frac{\exp\{h_r(\alpha+\xi+X'\beta)\}\partial h_r(\alpha+\xi+X'\beta)}{1+\sum_{r=1}^L \exp\{h_r(\alpha+\xi+X'\beta)\}}\right]}_{=:\tilde{h}_Y(\alpha+\xi+X'\beta)} \\
=&\int \sum_{l=1}^L \frac{\exp\{h_{l0}(\alpha_{i0}+\xi_{t0}+x'\beta_{i0})\}\tilde{h}_{0l}(\alpha_{i0}+\xi_{t0}+x'\beta_{i0})}{1+\sum_{r=1}^L \exp\{h_{r0}(\alpha_{i0}+\xi_{t0}+x'\beta_{i0})\}}\tilde{h}_l(\alpha+\xi+x'\beta)f(x|\alpha_{i0},\beta_{i0}|\xi_{t0})d(x,\alpha_{i0},\beta_{i0}) \\
+&\int \sum_{l=1}^L \frac{\exp\{h_{l0}(\alpha_{i0}+\xi_{t0}+x'\beta_{i0})\}\tilde{h}_l(\alpha+\xi+x'\beta)}{1+\sum_{r=1}^L \exp\{h_{r0}(\alpha_{i0}+\xi_{t0}+x'\beta_{i0})\}}\partial_\xi f_{0x}(x;\alpha_{i0},\beta_{i0},\xi_{t0})f_{0(\alpha,\beta)}(\alpha_{i0},\beta_{i0})d(x,\alpha_{i0},\beta_{i0}).
\end{aligned}
$$

Now using Conditions 1(d) and 2(c), we obtain $\left|\tilde{h}_l(\alpha+\xi+X'\beta)\right| \leq M_3$ and

$$
\left|\partial^2_{\xi_{t0}\xi}\mathbb{E}_{\cdot|\xi_{t0}}\left[\log\left(\frac{\exp\{h_Y(\alpha+\xi+X'\beta)\}}{1+\sum_{y=1}^L \exp\{h_y(\alpha+\xi+X'\beta)\}}\right)\right]\right| \leq M_3^2+M_3C_f.
$$

Then, $\left|\frac{\partial^2 \mathcal{L}_t(\xi;\{\alpha_i,\beta_i\}_{i=1}^N,h)}{\partial \xi_{t0}\partial \xi}\right| \leq M_3^2+M_3C_f$ and $\left|\frac{\partial \tilde{\xi}_t^*\left(\{\alpha_i,\beta_i\}_{i=1}^N,h\right)}{\partial \xi_{t0}}\right| \leq (M_3^2+M_3C_f)\lambda^{-1}$. For any (t,s) with $|\xi_{t0}-\xi_{s0}| \leq \delta$, we then have $\left|\tilde{\xi}_t^*\left(\{\alpha_i,\beta_i\}_{i=1}^N,h\right)-\tilde{\xi}_s^*\left(\{\alpha_i,\beta_i\}_{i=1}^N,h\right)\right| \leq (M_3^2+M_3C_f)\lambda^{-1}\delta$.

Given any $\varepsilon,\eta>0$, we can choose $T_{\varepsilon,\eta}$, and $N_{\varepsilon,\eta}$ such that

$$
\Pr\left(\underbrace{\sup_{(\alpha_i,\beta_i)\in\Omega_{\alpha,\beta},1\leq i\leq N;h\in\bar{\Theta}_h;1\leq t\leq T}\left|\tilde{\xi}\left((x_{it},y_{it})_{i=1}^N;\{\alpha_i,\beta_i\}_{i=1}^N,h\right)-\tilde{\xi}_t^*\left(\{\alpha_i,\beta_i\}_{i=1}^N,h\right)\right|}_{=:(\Delta_{NT}(\varepsilon,\eta)-\varepsilon/3)/2} > \frac{\varepsilon}{3}\right) < \eta.
$$

Then, for any (t,s) with $|\xi_{t0}-\xi_{s0}|<\delta_\varepsilon := \varepsilon/(3(M_3^2+M_3C_f)\lambda^{-1})$, we have:

$$
\begin{aligned}
&|\tilde{\xi}\left((x_{it},y_{it})_{i=1}^N;\{\alpha_i,\beta_i\}_{i=1}^N,h\right)-\tilde{\xi}\left((x_{is},y_{is})_{i=1}^N;\{\alpha_i,\beta_i\}_{i=1}^N,h\right)| \\
&\quad\leq 2\times(\Delta_{NT}(\varepsilon,\eta)-\varepsilon/3)/2+(M_3^2+M_3C_f)\lambda^{-1}\delta_\varepsilon \\
&\quad< \Delta_{NT}(\varepsilon,\eta).
\end{aligned}
$$

Note that $\Pr(\Delta_{NT}(\varepsilon,\eta)>\varepsilon)<\eta$.

Proof of Theorem 3. First, we construct a compact subset of $\mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta}))$, denoted by Θ_{G_1}, which $\hat{G}_1^{NT}$ and id_1 belong to. The construction of such a set of $\mathcal{C}(\Omega_\xi, \mathscr{P}(\Omega_\xi))$, denoted by Θ_{G_2}, is similar.

Note that for any $(\alpha,\beta) \in \Omega_{\alpha,\beta}$ and $\|(\alpha',\beta') - (\alpha,\beta)\| < \varepsilon$, we have $\|\mathrm{id}_1(\cdot|\alpha',\beta') - \mathrm{id}_1(\cdot|\alpha,\beta)\|_P = \|(\alpha',\beta') - (\alpha,\beta)\| < \varepsilon$. Let Θ_{G_1} be the subset of $\mathcal{C}(\Omega_{\alpha,\beta}, \mathscr{P}(\Omega_{\alpha,\beta}))$ that includes id_1 and is pointwise equicontinuous: at each $(\alpha,\beta) \in \Omega_{\alpha,\beta}$, for any $\varepsilon > 0$, there exists $\delta_\varepsilon(\alpha,\beta) < \min\{\delta_{\varepsilon/3}/3, \varepsilon\}$ such that

$$\sup_{G_1 \in \Theta_{G_1}, \|(\alpha',\beta')-(\alpha,\beta)\|_P < \delta_\varepsilon(\alpha,\beta)} \|G_1(\cdot|\alpha',\beta') - G_1(\cdot|\alpha,\beta)\|_P < \varepsilon \tag{26}$$

where δ_ε is defined in Assumption 7(b). Note that for any $(\alpha,\beta) \in \Omega_{\alpha,\beta}$, $\{G_1(\cdot|\alpha,\beta) : G_1 \in \Theta_{G_1}\}$ is a subset of the compact set $\mathscr{P}(\Omega_{\alpha,\beta})$. As a result, $\{G_1(\cdot|\alpha,\beta) : G_1 \in \Theta_{G_1}\}$ is relatively compact for any $(\alpha,\beta) \in \Omega_{\alpha,\beta}$. Then, according to Arzelá-Ascoli theorem, Θ_{G_1} is precompact and therefore its closure is compact in the metric $\|\cdot\|_1$. To simply notation, we use Θ_{G_1} to denote its closure. The construction of Θ_{G_2} is analogous.

We now prove the following proposition.

Proposition 1. *$\hat{G}_1^{NT} \in \Theta_{G_1}$ and $\hat{G}_2^{NT} \in \Theta_{G_2}$ asymptotically wp1.*

Proof. We prove $\hat{G}_1^{NT} \in \Theta_{G_1}$. The proof of $\hat{G}_2^{T} \in \Theta_{G_2}$ is similar. Define

$$\lambda^i(\alpha,\beta) := \prod_{r \neq i, r \in \mathcal{S}_N(\alpha,\beta)} \|(\alpha,\beta) - (\alpha_{r0},\beta_{r0})\|^2.$$

We will use the following lemma. The proof is in Online Appendix F.

Lemma 3. *Suppose that Assumption 4 holds. Then,*

$$\sup_{(\alpha,\beta)\in\Omega_{\alpha,\beta}} \min_{1\le i\le N} \|(\alpha,\beta) - (\alpha_{i0},\beta_{i0})\| \le \frac{2\ln N}{cC_2 N}$$

and

$$\sup_{\xi\in\Omega_\xi} \min_{1\le t\le T} |\xi - \xi_{t0}| \le \frac{1}{cT^{\frac{1}{3}}}$$

asymptotically wp1, where $C_2 = \frac{\pi^{3/2}}{\Gamma(5/2)}$ and μ is defined in Assumption 4.

Remark 1. *Lemma 3 implies that for any (α,β), when N is large enough so $\frac{\ln N}{\sqrt{N}} > \frac{2\ln N}{cC_2 N}$, there exists at least one i such that $(\alpha_{i0},\beta_{i0}) \in \mathcal{S}_N(\alpha,\beta)$. Similarly, when T is large enough, there exists at least one t such that $\xi_{t0} \in \mathcal{S}_N(\xi)$.*

Note that $(\alpha,\beta)=(\hat{\alpha}_i^{NT},\hat{\beta}_i^{NT})$ maximizes $\sum_{t=1}^{T}\log\frac{\exp\{\hat{h}_{y_{it}}^{NT}(\alpha+\hat{\xi}_t^{NT}+x_{it}'\beta)\}}{1+\sum_{y=1}^{L}\exp\{\hat{h}_{y}^{NT}(\alpha+\hat{\xi}_t^{NT}+x_{it}'\beta)\}}$. Then, according to Assumption 7(b), for any $\varepsilon/3,\eta>0$, there exist $N_{\varepsilon/3,\eta}$, $\delta_{\varepsilon/3}$ such that for any $N>N_{\varepsilon/3,\eta}$,

$$\sup_{1\leq i,r\leq N:\|(\alpha_{i0},\beta_{i0})-(\alpha_{r0},\beta_{r0})\|<\delta_{\varepsilon/3}}\|(\hat{\alpha}_i^{NT},\hat{\beta}_i^{NT})-(\hat{\alpha}_r^{NT},\hat{\beta}_r^{NT})\|<\varepsilon/3$$

holds on the event $\{\Delta_{NT}(\varepsilon/3,\eta)<\varepsilon/3\}$ whose probability is at least equal to $1-\eta$.

Fix a $(\alpha,\beta)\in\Omega_{\alpha,\beta}$. Now consider any $(\tilde{\alpha},\tilde{\beta})$ with $\|(\tilde{\alpha},\tilde{\beta})-(\alpha,\beta)\|<\delta_{\varepsilon/3}/3$. Define $i^*=\arg\min_{i\in\mathcal{S}_N(\alpha,\beta)}\|(\alpha-\alpha_{i0},\beta-\beta_{i0})\|$ and $\tilde{i}^*=\arg\min_{i\in\mathcal{S}_N(\tilde{\alpha},\tilde{\beta})}\|(\tilde{\alpha}-\alpha_{i0},\tilde{\beta}-\beta_{i0})\|$. Note that:

$$\begin{aligned}\|\hat{G}_1^{NT}(\cdot|\alpha,\beta)-\hat{G}_1^{NT}(\cdot|\tilde{\alpha},\tilde{\beta})\|_P &\leq \|\hat{G}_1^{NT}(\cdot|\alpha,\beta)-\hat{G}_1^{NT}(\cdot|\alpha_{i^*0},\beta_{i^*0})\|_P+\|\hat{G}_1^{NT}(\cdot|\tilde{\alpha},\tilde{\beta})-\hat{G}_1^{NT}(\cdot|\alpha_{\tilde{i}0},\beta_{\tilde{i}0})\|_P\\ &\quad+\|\hat{G}_1^{NT}(\cdot|\alpha_{i^*0},\beta_{i^*0})-\hat{G}_1^{NT}(\cdot|\alpha_{\tilde{i}0},\beta_{\tilde{i}0})\|_P\\ &=\|\hat{G}_1^{NT}(\cdot|\alpha,\beta)-\text{Dirac}_{(\hat{\alpha}_{i^*},\hat{\beta}_{i^*})}\|_P+\|\hat{G}_1^{NT}(\cdot|\tilde{\alpha},\tilde{\beta})-\text{Dirac}_{(\hat{\alpha}_{\tilde{i}},\hat{\beta}_{\tilde{i}})}\|_P\\ &\quad+\|\text{Dirac}_{(\hat{\alpha}_{i^*},\hat{\beta}_{i^*})}-\text{Dirac}_{(\hat{\alpha}_{\tilde{i}},\hat{\beta}_{\tilde{i}})}\|_P.\end{aligned}$$

Moreover,

$$\begin{aligned}\|\hat{G}_1^{NT}(\cdot|\alpha,\beta)-\text{Dirac}_{(\hat{\alpha}_{i^*},\hat{\beta}_{i^*})}\|_P &=\|\sum_{i\in\mathcal{S}_N(\alpha,\beta),i\neq i^*}\frac{\lambda^i(\alpha,\beta)}{\sum_{i\in\mathcal{S}_N(\alpha,\beta)}\lambda^i(\alpha,\beta)}\left(\text{Dirac}_{(\hat{\alpha}_i,\hat{\beta}_i)}-\text{Dirac}_{(\hat{\alpha}_{i^*},\hat{\beta}_{i^*})}\right)\|_P\\ &=\sup_{\|f\|_{BL}\leq 1}\left|\sum_{i\in\mathcal{S}_N(\alpha,\beta),i\neq i^*}\frac{\lambda^i(\alpha,\beta)}{\sum_{i\in\mathcal{S}_N(\alpha,\beta)}\lambda^i(\alpha,\beta)}\left(f(\hat{\alpha}_i,\hat{\beta}_i)-f(\hat{\alpha}_{i^*},\hat{\beta}_{i^*})\right)\right|\\ &\leq\sum_{i\in\mathcal{S}_N(\alpha,\beta),i\neq i^*}\frac{\|(\alpha-\alpha_{i0},\beta-\beta_{i0})\|^{-2}}{\sum_{i\in\mathcal{S}_N(\alpha,\beta)}\|(\alpha-\alpha_{i0},\beta-\beta_{i0})\|^{-2}}\|(\hat{\alpha}_i-\hat{\alpha}_{i^*},\hat{\beta}_i-\hat{\beta}_{i^*})\|.\end{aligned}$$

By the definition of $\mathcal{S}_N(\alpha,\beta)$, when N is large enough, $\|(\alpha_{i0}-\alpha_{i^*0},\beta_{i0}-\beta_{i^*0})\|\leq\|(\alpha-\alpha_{i^*0},\beta-\beta_{i^*0})\|+\|(\alpha_{i0}-\alpha,\beta_{i0}-\beta)\|\leq\frac{2\ln N}{\sqrt{N}}\leq\frac{2\delta_{\varepsilon/3}}{3}<\delta_{\varepsilon/3}$ for all $i\in\mathcal{S}_N(\alpha,\beta)$. As a result,

$$\|\hat{G}_1^{NT}(\cdot|\alpha,\beta)-\text{Dirac}_{(\hat{\alpha}_{i^*},\hat{\beta}_{i^*})}\|_P\leq\sum_{i\in\mathcal{S}_N(\alpha,\beta),i\neq i^*}\frac{\|(\alpha-\alpha_{i0},\beta-\beta_{i0})\|^{-2}}{\sum_{i\in\mathcal{S}_N(\alpha,\beta)}\|(\alpha-\alpha_{i0},\beta-\beta_{i0})\|^{-2}}\frac{\varepsilon}{3}<\frac{\varepsilon}{3}$$

holds on the event $\{\Delta_{NT}(\varepsilon/3,\eta)<\varepsilon/3\}$. Similarly,

$$\|\hat{G}_1^{NT}(\cdot|\tilde{\alpha},\tilde{\beta})-\text{Dirac}_{(\hat{\alpha}_{\tilde{i}^*},\hat{\beta}_{\tilde{i}^*})}\|_P<\frac{\varepsilon}{3}$$

holds on the same event. Moreover,

$$\|\text{Dirac}_{(\hat{\alpha}_{i^*},\hat{\beta}_{i^*})} - \text{Dirac}_{(\hat{\alpha}_{\tilde{i}},\hat{\beta}_{\tilde{i}})}\|_P = \|(\hat{\alpha}_{i^*},\hat{\beta}_{i^*}) - (\hat{\alpha}_{\tilde{i}^*},\hat{\beta}_{\tilde{i}^*})\|$$

and

$$\|(\alpha_{i^*0},\beta_{i^*0}) - (\alpha_{\tilde{i}^*0},\beta_{\tilde{i}^*0})\| \leq \frac{2\ln N}{\sqrt{N}} + \|(\tilde{\alpha},\tilde{\beta}) - (\alpha,\beta)\| < \delta_{\varepsilon/3}.$$

Then, $\|\text{Dirac}_{(\hat{\alpha}_{i^*},\hat{\beta}_{i^*})} - \text{Dirac}_{(\hat{\alpha}_{\tilde{i}},\hat{\beta}_{\tilde{i}})}\|_P < \varepsilon/3$ on the event $\{\Delta_{NT}(\varepsilon/3,\eta) < \varepsilon/3\}$. Consequently, $\|\hat{G}_1^{NT}(\cdot|\alpha,\beta) - \hat{G}_1^{NT}(\cdot|\tilde{\alpha},\tilde{\beta})\|_P < \varepsilon$ for any $\|(\tilde{\alpha},\tilde{\beta}) - (\alpha,\beta)\| < \delta_{\varepsilon/3}/3$ and therefore $\hat{G}_1^{NT} \in \Theta_{G_1}$ on the event $\{\Delta_{NT}(\varepsilon/3,\eta) < \varepsilon/3\}$ whose probability is at least equal to $1-\eta$. This conclusion holds for any $\eta > 0$. Proposition 1 is then proved. □

For any $h \in \bar{\Theta}_h$ and realizations $(y_{it})_{i,t}$, the following inequality holds asymptotically with probability 1 (wp1):

$$\begin{aligned}
&\left|\mathscr{L}_{NT}(\theta_{10}^{NT},h) - \mathscr{L}_{NT}(\theta_{10}^{NT},h_0)\right| \\
\leq &\frac{1}{NT}\sum_{i,t}|h_{y_{it}}(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0}) - h_{y_{it}0}(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0})| \\
+&\frac{1}{NT}\sum_{i,t}\left|\log\left(1+\sum_{y=1}^{L}\exp\{h_y(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0})\}\right) - \log\left(1+\sum_{y=1}^{L}\exp\{h_{y0}(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0})\}\right)\right| \\
\leq &\frac{1}{NT}\sum_{i,t}\left[|h_{y_{it}}(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0}) - h_{y_{it}0}(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0})| + \max_{1\leq y\leq L}\{|h_y(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0}) - h_{y0}(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0})|\}\right] \\
\leq &\frac{2}{NT}\sum_{i,t}\max_{1\leq y\leq L}|h_y(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0}) - h_{y0}(\alpha_{i0}+\xi_{t0}+x_{it}'\beta_{i0})| \\
\leq &2\|h-h_0\|_h.
\end{aligned} \tag{27}$$

We now prove Theorem 3. Denote by h_{l0}^{NT} the projection of h_{l0} on $\Theta_{h_l}^{NT}$ for $l = 1,\ldots,L$ and $h_0^{NT} = (h_{l0}^{NT})_{l=1}^{L}$. By the definition of $\hat{\theta}^{NT}$ in (8) and the uniform convergence in Assumption 7(a), we have for any $\varepsilon > 0$,

$$\mathscr{L}_{NT}^0(\hat{\theta}^{NT}) > \mathscr{L}_{NT}^0(\theta_{10}^{NT},h_0^{NT}) - \varepsilon/3 \tag{28}$$

asymptotically wp1. Moreover, by Assumption 7(c), we have

$$\begin{aligned}
\left|\mathscr{L}_{NT}^0(\theta_{10}^{NT},h_0) - \mathscr{L}^0(\text{id}_1,\text{id}_2,h_0;F_0)\right| &= \left|\mathscr{L}^0(\text{id}_1,\text{id}_2,h_0;F_0^{NT}) - \mathscr{L}^0(\text{id}_1,\text{id}_2,h_0;F_0)\right| \\
&\leq \varepsilon/3
\end{aligned} \tag{29}$$

asymptotically wp1. Besides, by (27) and Assumption 5 we have: when N and T

are large enough, $\|h_0^{NT} - h_0\|_h$ is so small that

$$|\mathscr{L}_{NT}^0(\theta_{10}^{NT}, h_0^{NT}) - \mathscr{L}_{NT}^0(\theta_{10}^{NT}, h_0)| < \varepsilon/3. \tag{30}$$

Then, combining (28)–(30), we have:

$$\mathscr{L}_{NT}^0(\hat{\theta}^{NT}) > \mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0) - \varepsilon \tag{31}$$

asymptotically wp1. Using the definition of $\hat{G}^{NT}$ in (10), we can express (31) as

$$\mathscr{L}_{NT}^0(\hat{\theta}^{NT}) = \mathscr{L}^0(\hat{G}^{NT}, \hat{h}^{NT}; F_0^{NT}) > \mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0) - \varepsilon.$$

Using Gibbs' inequality, we have

$$\mathscr{L}^0(\hat{G}^{NT}, \hat{h}^{NT}; F_0^{NT}) \leq \mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0^{NT}).$$

As a result,

$$\mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0^{NT}) \geq \mathscr{L}^0(\hat{G}^{NT}, \hat{h}^{NT}; F_0^{NT}) > \mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0) - \varepsilon. \tag{32}$$

hold asymptotically wp1.

Using the compactness of $\Theta_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h$ and Assumption 6, we have: when ε is small enough, we can find $\delta(\varepsilon) > 0$ such that

$$\mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0) > \sup_{\max\{\|G_1 - \mathrm{id}_1\|_1, \|G_2 - \mathrm{id}_2\|_2, \|h - h_0\|_h\} \geq \delta(\varepsilon), (G,h) \in \Theta_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h} \mathscr{L}^0(G, h; F_0) + 3\varepsilon. \tag{33}$$

Using Assumption 7(c), we then have: for $\varepsilon/2$,

$$\left|\mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0^{NT}) - \mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0)\right| \leq \varepsilon/2 \tag{34}$$

and

$$\sup_{\max\{\|G_1 - \mathrm{id}_1\|_1, \|G_2 - \mathrm{id}_2\|_2, \|h - h_0\|_h\} \geq \delta(\varepsilon), (G,h) \in \Theta_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h} \left|\mathscr{L}^0(G, h; F_0^{NT}) - \mathscr{L}^0(G, h; F_0)\right| \leq \varepsilon/2 \tag{35}$$

asymptotically wp1. Then, using (34) and (32), we obtain:

$$\mathscr{L}^0(\hat{G}^{NT}, \hat{h}^{NT}; F_0^{NT}) > \mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0^{NT}) - 3\varepsilon/2.$$

Moreover,

$$
\begin{aligned}
&\mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0^{NT}) \\
&\geq \mathscr{L}^0(\mathrm{id}_1, \mathrm{id}_2, h_0; F_0) - \varepsilon/2 \ \ \text{(by (34))} \\
&\geq \sup_{\max\{\|G_1-\mathrm{id}_1\|_1, \|G_2-\mathrm{id}_2\|_2, \|h-h_0\|_h\} \geq \delta(\varepsilon), (G,h) \in \Theta_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h} \mathscr{L}^0(G, h; F_0) + 5\varepsilon/2 \ \ \text{(by (33))} \\
&= \sup_{\max\{\|G_1-\mathrm{id}_1\|_1, \|G_2-\mathrm{id}_2\|_2, \|h-h_0\|_h\} \geq \delta(\varepsilon), (G,h) \in \bar{\Theta}_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h} \left(\mathscr{L}^0(G, h; F_0) - \mathscr{L}^0(G, h; F_0^{NT}) + \mathscr{L}^0(G, h; F_0^{NT})\right) + 5\varepsilon/2 \\
&\geq \sup_{\max\{\|G_1-\mathrm{id}_1\|_1, \|G_2-\mathrm{id}_2\|_2, \|h-h_0\|_h\} \geq \delta(\varepsilon), (G,h) \in \Theta_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h} \mathscr{L}^0(G, h; F_0^{NT}) \\
&\quad - \sup_{\max\{\|G_1-\mathrm{id}_1\|_1, \|G_2-\mathrm{id}_2\|_2, \|h-h_0\|_h\} \geq \delta(\varepsilon), (G,h) \in \Theta_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h} \left|\mathscr{L}^0(G, h; F_0) - \mathscr{L}^0(G, h; F_0^{NT})\right| + 5\varepsilon/2 \\
&\geq \sup_{\max\{\|G_1-\mathrm{id}_1\|_1, \|G_2-\mathrm{id}_2\|_2, \|h-h_0\|_h\} \geq \delta(\varepsilon), (G,h) \in \Theta_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h} \mathscr{L}^0(G, h; F_0^{NT}) + 2\varepsilon \ \ \text{(by (35))}.
\end{aligned}
$$

holds asymptotically wp1. Consequently,

$$
\mathscr{L}^0(\hat{G}^{NT}, \hat{h}^{NT}; F_0^{NT}) > \sup_{\max\{\|G_1-\mathrm{id}_1\|_1, \|G_2-\mathrm{id}_2\|_2, \|h-h_0\|_h\} > \delta(\varepsilon), (G,h) \in \Theta_{G_1} \times \Theta_{G_2} \times \bar{\Theta}_h} \mathscr{L}^0(G, h; F_0^{NT}) + \varepsilon/2.
$$

Then, we have $\max\{\|G_1 - \mathrm{id}_1\|_1, \|G_2 - \mathrm{id}_2\|_2, \|h - h_0\|_h\} < \delta(\varepsilon)$ asymptotically wp1. This implies

$$
\|\hat{G}_1^{NT} - \mathrm{id}_1\|_1 = \sup_{(\alpha,\beta) \in \Omega_{\alpha,\beta}} \|\hat{G}_1^{NT}(\cdot|\alpha, \beta) - \mathrm{Dirac}_{(\alpha,\beta)}(\cdot)\| < \delta(\varepsilon),
$$

$$
\|\hat{G}_2^{NT} - \mathrm{id}_2\|_2 = \sup_{\xi \in \Omega_\xi} \|\hat{G}_2^{NT}(\cdot|\xi) - \mathrm{Dirac}_{\xi}(\cdot)\| < \delta(\varepsilon),
$$

and

$$
\|\hat{h}^{NT} - h_0\|_h < \delta(\varepsilon).
$$

The proof is completed.

Online Appendix

E Extensions

In this appendix, we discuss the extensions of Theorems 1 and 2 to models with heterogeneous slopes across time and with multinomial outcomes.

Heterogeneous slope across time. The case of heterogeneous slope across time mirrors the case of heterogeneous slope across individuals in terms of individual and time dimensions. We can apply the argument of compensating variable across time periods to obtain the identification results. To start with, we modify Assumption 1.

Assumption 1′ (Model)**.**

(a) Single index and two-way fixed effects: for all (i,t)*,*

$$\Pr\left(Y_{it} = y | (X_{is})_{s\leq t}, \mathcal{F}\right) = g(y; X_{it}'\beta_t + \alpha_i + \xi_t),$$

with, almost surely, $\sup_i \|\beta_t\| \leq C_\beta < \infty$*. Moreover,* g *is unknown.*

(b) Monotonicity and smoothness: there exists $\bar{y} \in \mathcal{Y}$ *such that the function* $v \mapsto g(\bar{y}; v)$ *is strictly increasing and L-Lipschitz.*

(c) Cross-section independence and weak serial dependence:

1. *Conditional on* $\mathcal{F}$*,* $\{(Y_{it}, X_{it}') : t = 1, 2, \ldots\}$ *is independent across* i*.*
2. *Let* $\mu > 1$*. Conditional on* $\mathcal{F}$*, for each* i*,* $\{(Y_{it}, X_{it}') : t = 1, 2, \ldots\}$ *is* α*-mixing with mixing coefficient satisfying* $\sup_i a_i(m) = O(m^{-\mu})$ *as* $m \to \infty$*, where*

$$a_i(m) := \sup_t \sup_{A \in \mathcal{A}_t^i, B \in \mathcal{B}_{t+m}^i} \left|\Pr(A \cap B) - \Pr(A)\Pr(B)\right|,$$

and for $Z_{it} := (Y_{it}, X_{it}')$*,* $\mathcal{A}_t^i$ *is the* σ*-field generated by* $(Z_{it}, Z_{it-1}, \ldots)$*, and* $\mathcal{B}_t^1$ *is the* σ*-field generated by* $(Z_{it}, Z_{it+1}, \ldots)$*.*

(d) Conditional on $\mathcal{F}$*,* X_{it} *has density* p_{it} *with respect to the Lebesgue measure on* $\mathbf{R}^K$ *such that* $p_{it}(x) \leq p_{max} < \infty$ *for all* (i,t) *and* $x \in \mathbf{R}^K$*.*

(e) *Let K denote a bounded kernel function. For all strictly monotonic functions $f : \mathbf{R} \to (0,1)$ and $x = (x^{(1)}, x^{(2)}) \in \mathcal{X}$, almost surely, there exists a constant $c_{f,x^{(2)}}$ nontrivially depending on f such that, for all t,*

$$\frac{\frac{1}{h_N N}\sum_{i=1}^{N} K\left(\frac{X_{it}-x}{h_N}\right) f(\alpha_i)}{\frac{1}{h_N N}\sum_{i=1}^{N} K\left(\frac{X_{it}-x}{h_N}\right)} \to c_{f,x^{(2)}} \text{ as } N \to \infty.$$

Define a compensating variable:

$$z_{t\to t'}(x^{(1)}; x^{(2)}) = [x^{(1)}\beta_{t'}^{(1)} + x^{(2)}(\beta_{t'}^{(2)} - \beta_t^{(2)}) + \xi_{t'} - \xi_t]/\beta_t^{(1)}. \tag{36}$$

Intuitively, $z_{t\to t'}(x^{(1)}; x^{(2)})$ is the needed value of $x^{(1)}$ for individual i with $x^{(2)}$ at time t to make her indices at time t' and t equal: $\alpha_i + \xi_t + \beta_t^{(1)} z_{t\to t'}(x^{(1)}; x^{(2)}) + \beta_t^{(2)} x^{(2)} = \alpha_i + \xi_{t'} + \beta_{t'}^{(1)} x^{(1)} + \beta_{t'}^{(2)} x^{(2)}$.

Definition 1′. *Time period t is compensable at $(x^{(1)}, x^{(2)}) \in \mathcal{X}^{t'}$ by time period t if and only if $(z_{t\to t'}(x^{(1)}; x^{(2)}), x^{(2)}) \in \mathcal{X}^t$.*[12]

With Definition 1′, we can then define a compensating network of time periods and achieve the relative identification (along the time dimension) similar to Theorem 1. Under conditions analogous to Assumption 3, we can further achieve point identification similarly to Theorem 2.

Multinomial outcomes. Consider a model with multinomial outcomes: the probability of individual i from choosing $y_{it} \in \{1, \ldots, J\}$ at time t is

$$\Pr(y_{it} = j \mid (\alpha_{ij}, \xi_{sj}, \beta_{ij}, x_{isj})_{j=1,\ldots,J}^{s=1,\ldots,t}) = g_j(v_{it}), \tag{37}$$

where $v_{it} = (v_{itj})_{j=1}^J$ with $v_{itj} = \alpha_{ij} + \xi_{tj} + x'_{itj}\beta_{ij}$, $\sum_{j=1}^J g_j(v_{it}) < 1$, and J is known. The residual probability, $g_0(v_{it}) = 1 - \sum_{j=1}^J g_j(v_{it})$, is usually defined as the probability of choosing the outside option. Model (37) is a common setting in empirical research such as demand estimation (Berry et al., 2013; Dubois et al., 2020). Define $g(v_{it}) = (g_j(v_{it}))_{j=1}^J$, $\alpha_i = (\alpha_{ij})_{j=1}^J$, $\beta_i = (\beta_{ij})_{j=1}^J$, $\xi_t = (\xi_{tj})_{j=1}^J$.

To extend the identification results to (37), we first replace the monotonicity in Assumption 1(b) by the invertibility of $g(v)$, i.e., the mapping $g(\cdot)$ is a bijection. The following assumption provides a set of sufficient conditions for the invertibility.

[12] We assume that $\mathcal{X}^t$ does not depend on i to simplify the exposition. This holds, e.g., if $\{x_{it}\}_{i\geq 2}$ is i.i.d. conditional on (ξ_t, β_t) for all $t \in \mathbf{T}$.

Assumption 1$(b)'$**.** *The mapping g satisfies the following conditions:*

- *The support of $(v_{it1}, \ldots, v_{itJ})$, $\mathcal{V}$, is a Cartesian product.*
- *(Weak substitutes) $g_j(v)$ is weakly decreasing in v_k for all $j = 1, \ldots, J$ and $k \notin \{0, j\}$.*
- *(Connected strict substitution) For all $v \in \mathcal{V}$ and any nonempty subset of $\{1, \ldots, J\}$, $\mathcal{K}$, there exists $k \in \mathcal{K}$ and $l \notin \mathcal{K}$ such that g_l is strictly decreasing in v_k.*

Assumption 1$(b)'$ is a multi-index version of Assumption 1(b). It is motivated by sufficient conditions for the invertibility of demand by Berry et al. (2013) and usually satisfied in the setting of discrete-choice random utility models with separably additive index and idiosyncratic error in indirect utility. This assumption implies that g is a bijection from $\mathcal{V}$ to $g(\mathcal{V})$. Moreover, it implies that the aggregated choice probability function, i.e., the integral of $g(v_{it})$ over ξ_t for a given i, satisfies Assumption 1$(b)'$ and is therefore a bijection. Both bijection properties enable to apply the argument of compensating variable as in the single-index case. As argued in Berry et al. (2013), Assumption 1$(b)'$ is convenient in practice due to its Cartesian support requirement and it applies even when g may not be not differentiable. This contrasts other arguments such as those by Gale and Nikaido (1965) which require rectangular support condition and differentiability of g. In contrast, due to the weak substitutes requirement in Assumption 1$(b)'$, the derivative of g_j (if differentiable) with respect to v_k is restricted to be nonpositive for all $k \neq j$. As an alternative, one can require $g_j(\cdot)$ to be strictly increasing with respect to index v_j for all $j = 1, \ldots, J$ and the mapping g to have strictly diagonally dominant Jacobian, which will also imply the bijection properties we need to apply the argument of compensating variable but allows for positive cross derivatives in g. Second, analogously to the single-index case, we define a compensating vector of dimension J:

$$z_{i\to i'}(x^{(1)}; x^{(2)}) = (z_{i\to i',j}(x^{(1)}; x^{(2)}))_{j=1}^{J} = \left([\alpha_{i'j} - \alpha_{ij} + x_j^{(1)}\beta_{i'j}^{(1)} + x_j^{(2)}(\beta_{i'j}^{(2)} - \beta_{ij}^{(2)})]/\beta_{ij}^{(1)}\right)_{j=1}^{J}$$

$z_{i\to i'}(x^{(1)}; x^{(2)})$ is the needed value of $x^{(1)}$ for individual i with $x^{(2)}$ to make her and i's indices equal: for $j = 1, \ldots, J$,

$$\alpha_{ij} + \xi_{tj} + \beta_{ij}^{(1)} z_{i\to i',j}(x^{(1)}; x^{(2)}) + \beta_{ij}^{(2)} x_j^{(2)} = \alpha_{i'j} + \xi_{tj} + \beta_{i'j}^{(1)} x_j^{(1)} + \beta_{i'j}^{(2)} x_j^{(2)}.$$

The definition of compensation and compensating network should be accordingly modified. In particular, the rank condition in the definition of compensating network should now hold for all $j = 1, ..., J$: individual i is compensable by individual i' at least at $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_i$ for $k = 1, 2, 3$ with

$$\begin{bmatrix} 1 & x_j^{(1)1} & x_j^{(2)1} \\ 1 & x_j^{(1)2} & x_j^{(2)2} \\ 1 & x_j^{(1)3} & x_j^{(2)3} \end{bmatrix}$$

being nonsingular for $j = 1, \ldots, J$. With these modifications, one can readily apply the arguments in the proofs of Theorems 1 and 2.

F Proofs of Lemmas

Proof of Lemma 1. Define $\Delta_{it}^1 := \frac{M_{it}^{NT} K\left(\frac{X_{it}-x}{h_T}\right)\mathbf{1}\{Y_{it}=\overline{y}\}}{h_T \sum_{t=1}^{\infty} M_{it}^{NT}}$ and $\Delta_{it}^2 := \frac{M_{it}^{NT} K\left(\frac{X_{it}-x}{h_T}\right) g(\overline{y}; X_{it}'\beta_i + \alpha_i + \xi_t)}{h_T \sum_{t=1}^{\infty} M_{it}^{NT}}$.

First Part. We show:

$$\sum_t (\Delta_{it}^1 - \Delta_{it}^2) = o_p(1). \tag{38}$$

Let $\mathbb{E}_{\mathcal{F}}$ (resp. $\text{Var}_{\mathcal{F}}$) denotes the expectation (resp. variance) with respect to the distribution of the data conditional on the unobserved effects $\mathcal{F}$. For all t, $\mathbb{E}_{\mathcal{F}}\left(\Delta_{it}^1 - \Delta_{it}^2 | X_i^t, M^{NT}\right) = 0$.

$$\begin{aligned} &\text{Var}_{\mathcal{F}}\left(\sum_{t=1}^{\infty} \Delta_{it}^1 - \Delta_{it}^2\right) \\ &\quad = \mathbb{E}_{\mathcal{F}}\left[\sum_{t=1}^{\infty} \text{Var}_{\mathcal{F}}\left(\Delta_{it}^1 - \Delta_{it}^2 | X_i^t, M^{NT}\right)\right] + 2\sum_{s<t} \text{Cov}_{\mathcal{F}}\left(\Delta_{is}^1 - \Delta_{is}^2, \Delta_{it}^1 - \Delta_{it}^2\right). \end{aligned} \tag{39}$$

Focus on the first term. We have

$$\text{Var}_{\mathcal{F}}\left(\Delta_{it}^1 - \Delta_{it}^2 | X_i^t, M^{NT}\right) = \frac{M_{it}^{NT} K^2\left(\frac{X_{it}-x}{h_T}\right)}{h_T^2 \left(\sum_{t=1}^{\infty} M_{it}^{NT}\right)^2} \delta_{it}(X_{it})(1 - \delta_{it}(X_{it})),$$

where $\delta_{it}(X_{it}) := g(\overline{y}; X_{it}'\beta_i + \alpha_i + \xi_t)$. Because $0 \leq \delta_{it}(X_{it})(1 - \delta_{it}(X_{it})) \leq 1$,

$\sum_{t=1}^{\infty} M_{it}^{NT} = T$, $\left(\sum_t M_{it}^{NT}\right)^2 = T^2$, and

$$\mathbb{E}_{\mathcal{F}}\left[K^2\left(\frac{X_{it}-x}{h_T}\right)\middle| M^{NT}\right] \leq h_T C,$$

where $C := p_{\max}\int K^2(u)du < \infty$, we have

$$\mathbb{E}_{\mathcal{F}}\left[\sum_{t=1}^{\infty}\mathrm{Var}_{\mathcal{F}}\left(\Delta_{it}^1 - \Delta_{it}^2 | X_i^t, M^{NT}\right)\middle| M^{NT}\right] \leq \frac{h_T C\sum_{t=1}^{\infty} M_{it}^{NT}}{T^2 h_T^2} \leq \frac{C}{T h_T}.$$

Hence, as N, T tend to infinity,

$$\mathbb{E}_{\mathcal{F}}\left[\sum_{t=1}^{\infty}\mathrm{Var}_{\mathcal{F}}\left(\Delta_{it}^1 - \Delta_{it}^2 | X_i^t, M^{NT}\right)\right] \to 0$$

We now turn to the second term in (39). For all (s,t) such that $s < t$, we have

$$\mathbb{E}_{\mathcal{F}}\left[\Delta_{is}^1 - \Delta_{is}^2 | X_i^s, M^{NT}\right] = \mathbb{E}_{\mathcal{F}}\left[\Delta_{it}^1 - \Delta_{it}^2 | X_i^s, M^{NT}\right] = 0.$$

Hence, it suffices to show that, almost surely,

$$\mathrm{Cov}_{\mathcal{F}}\left(\Delta_{is}^1 - \Delta_{is}^2, \Delta_{it}^1 - \Delta_{it}^2 | M^{NT}\right) \to 0.$$

Let $\zeta_{it} := \frac{1}{h_T}K\left(\frac{X_{it}-x}{h_T}\right)\left[\mathbf{1}\{Y_{it} = \overline{y}\} - g(y; X_{it}'\beta_i + \alpha_i + \xi_t)\right]$. We have

$$\begin{aligned}
&\mathrm{Cov}_{\mathcal{F}}\left(\Delta_{is}^1 - \Delta_{is}^2, \Delta_{it}^1 - \Delta_{it}^2 | M^{NT}\right) \\
&= \frac{M_{is}^{NT} M_{it}^{NT}}{h_T^2 T^2}\mathrm{Cov}_{\mathcal{F}}\left(K\left(\frac{X_{is}-x}{h_T}\right)\mathbf{1}\{Y_{is} = \overline{y}\}, K\left(\frac{X_{it}-x}{h_T}\right)\mathbf{1}\{Y_{it} = \overline{y}\}\right) \\
&= \frac{M_{is}^{NT} M_{it}^{NT}}{T^2}\mathrm{Cov}_{\mathcal{F}}\left(\zeta_{is}, \zeta_{it} | M^{NT}\right).
\end{aligned}$$

Conditional on $\mathcal{F}, M^{NT}$, $\{\zeta_{it} : t = 1, 2, \ldots\}$ is α-mixing. Moreover, for all t,

$$\begin{aligned}
\mathbb{E}_{\mathcal{F}}\left(|\zeta_{it}|^{2+\delta} | M^{NT}\right) &= \mathbb{E}_{\mathcal{F}}\left(\left|\frac{1}{h_T}K\left(\frac{X_{it}-x}{h_T}\right)\mathbf{1}\{Y_{it} = \overline{y}\}\right|^{2+\delta}\right) \\
&\leq \mathbb{E}_{\mathcal{F}}\left(\left|\frac{1}{h_T}K\left(\frac{X_{it}-x}{h_T}\right)\right|^{2+\delta}\right) \\
&\leq C' < \infty.
\end{aligned}$$

By Proposition 2.5 in Fan and Yao (2003), there exists a positive constant $C'' < \infty$

such that

$$\mathrm{Cov}_{\mathcal{F}}\left(\zeta_{it},\zeta_{is}|M^{NT}\right)\leq C''a_i(|t-s|)^{\delta/(2+\delta)}.$$

Hence, by majoring the sum of covariances by the sum of covariances for the closest possible time sequence (all successive periods), we have

$$\mathbb{E}_{\mathcal{F}}\left[\sum_{s<t}\frac{M_{is}^{NT}M_{it}^{NT}}{T^2}\mathrm{Cov}_{\mathcal{F}}\left(\zeta_{it},\zeta_{is}|M^{NT}\right)\right]\leq C''\frac{1}{T^2}\sum_{1\leq s<t\leq T}a_i(|t-s|)^{\delta/(2+\delta)}\to 0.$$

This shows (38).

Second part. Now given Equation (38), and because

$$\Delta_i^3:=\frac{1}{h_T\sum_{t=1}^{\infty}M_{it}^{NT}}\sum_{t=1}^{\infty}M_{it}^{NT}K\left(\frac{X_{it}-x}{h_T}\right)=O_p(1),$$

we have

$$\sum_{t=1}\frac{\Delta_{it}^1-\Delta_{it}^2}{\Delta_i^3}=o_p(1).$$

It remains to show that

$$\frac{\sum_t\Delta_{it}^2}{\Delta_i^3}=\Gamma_i(y,x)+o_p(1). \tag{40}$$

Then, by the continuous mapping theorem, we have

$$\widehat{\Gamma}_i(y,x)-\Gamma_i(y,x)=\sum_{t=1}\frac{\Delta_{it}^1-\Delta_{it}^2}{\Delta_i^3}+\sum_{t=1}\frac{\Delta_{it}^2}{\Delta_i^3}-\Gamma_i(y,x)=o_p(1).$$

Consider the decomposition

$$\begin{aligned}\frac{\sum_t\Delta_{it}^2}{\Delta_i^3}=&\frac{\frac{1}{h_T\sum_{t=1}^{\infty}M_{it}^{NT}}\sum_{t=1}^{\infty}M_{it}^{NT}K\left(\frac{X_{it}-x}{h_T}\right)\left[g(\overline{y},X_{it}'\beta_i+\alpha_i+\xi_t)-g(\overline{y},x'\beta_i+\alpha_i+\xi_t)\right]}{\frac{1}{h_T\sum_{t=1}^{\infty}M_{it}^{NT}}\sum_{t=1}^{\infty}M_{it}^{NT}K\left(\frac{X_{it}-x}{h_T}\right)}\\&+\frac{\frac{1}{h_T\sum_{t=1}^{\infty}M_{it}^{NT}}\sum_{t=1}^{\infty}M_{it}^{NT}K\left(\frac{X_{it}-x}{h_T}\right)g(\overline{y},x'\beta_i+\alpha_i+\xi_t)}{\frac{1}{h_T\sum_{t=1}^{\infty}M_{it}^{NT}}\sum_{t=1}^{\infty}M_{it}^{NT}K\left(\frac{X_{it}-x}{h_T}\right)}.\end{aligned}$$

By Assumptions 1(a),(b), and (e), the second term converges almost surely to $c_{g(\overline{y},x'\beta_i+\alpha_i+\cdot),x^{(2)}}$, and $\Gamma_i(\overline{y},x):=c_{g(\overline{y},x'\beta_i+\alpha_i+\cdot),x^{(2)}}$ is strictly monotonic in $x'\beta_i+\alpha_i$.

Let us show that the first term is $o_p(1)$.

$$A_{NT} := \left| \frac{1}{h_T \sum_{t=1}^{\infty} M_{it}^{NT}} \sum_{t=1}^{\infty} M_{it}^{NT} K\left(\frac{X_{it}-x}{h_T}\right) [g(\overline{y}, X_{it}'\beta_i + \alpha_i + \xi_t) - g(\overline{y}, x'\beta_i + \alpha_i + \xi_t)] \right|$$
$$\leq \frac{L}{h_T T} \sum_{t=1}^{\infty} M_{it}^{NT} \left| K\left(\frac{X_{it}-x}{h_T}\right) \right| |(X_{it}-x)'\beta_i|$$
$$\leq \frac{LC_\beta}{h_T T} \sum_{t=1}^{\infty} M_{it}^{NT} \left| K\left(\frac{X_{it}-x}{h_T}\right) \right| \|X_{it}-x\|.$$

Then, by similar argument as before, there exists a positive constant $:= C''' < \infty$ such that uniformly over N, T, i, t,

$$\mathbb{E}_{\mathcal{F}} \left[\left| K\left(\frac{X_{it}-x}{h_T}\right) \right| \|X_{it}-x\| \, |M^{NT} \right] \leq := C''' h_T^2.$$

Then,

$$\mathbb{E}_{\mathcal{F}} \left[A_{NT} | M^{NT} \right] \leq \frac{LC_\beta := C'''}{h_T T} \sum_{t=1}^{\infty} M_{it}^{NT} h_T^2 = LC_\beta := C''' h_T = o_p(1).$$

Proof of Lemma 2. ($\Longrightarrow$) Suppose $i \longleftrightarrow i'$ in $\mathcal{G}^\infty$ and, without loss of generality, $i \to i'$. By the first part of Lemma 1, for some $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_{i'}$ for $k = 1, 2, 3$ with

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix}$$

being nonsingular, we have

$$\Gamma_i(\overline{y}, (z_{i\to i'}(x^{(1)k}; x^{(2)k}), x^{(2)k})) = \Gamma_{i'}(\overline{y}, (x^{(1)k}, x^{(2)k}))$$

with $z_{i\to i'}(x^{(1)k}; x^{(2)k}), x^{(2)k}) \in \mathcal{X}_i$. Hence, letting $\tilde{x}^{(1)k} := z_{i\to i'}(x^{(1)k}; x^{(2)k}), x^{(2)k})$, there exist some $(\tilde{x}^{(1)k}, \tilde{x}^{(2)k}) \in \mathcal{X}_i$, $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_{i'}$ for $k = 1, 2, 3$ such that

$$\Gamma_i(\overline{y}, (\tilde{x}^{(1)k}, \tilde{x}^{(2)k})) = \Gamma_{i'}(\overline{y}, (x^{(1)k}, x^{(2)k})),$$

with

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix}$$

being nonsingular. When $i' \to i$, we obtain the same equality with

$$\begin{bmatrix} 1 & \tilde{x}^{(1)1} & \tilde{x}^{(2)1} \\ 1 & \tilde{x}^{(1)2} & \tilde{x}^{(2)2} \\ 1 & \tilde{x}^{(1)3} & \tilde{x}^{(2)3} \end{bmatrix}$$

being nonsingular.

($\Longleftarrow$) Suppose that there exist $(\tilde{x}^{(1)k}, \tilde{x}^{(2)k}) \in \mathcal{X}_i$, $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_{i'}$ for $k = 1, 2, 3$ such that

$$\Gamma_i(\overline{y}, (\tilde{x}^{(1)k}, \tilde{x}^{(2)k})) = \Gamma_{i'}(\overline{y}, (x^{(1)k}, x^{(2)k})),$$

with either

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix} \text{ or } \begin{bmatrix} 1 & \tilde{x}^{(1)1} & \tilde{x}^{(2)1} \\ 1 & \tilde{x}^{(1)2} & \tilde{x}^{(2)2} \\ 1 & \tilde{x}^{(1)3} & \tilde{x}^{(2)3} \end{bmatrix}$$

being nonsingular. By the second part of Lemma 1, $\tilde{x}^{(1)k} = z_{i \to i'}(x^{(1)k}; x^{(2)k}), x^{(2)k})$ so that $i \to i'$, or $x^{(1)k} = z_{i' \to i}(\tilde{x}^{(1)k}; \tilde{x}^{(2)k}), \tilde{x}^{(2)k})$ so that $i' \to i$. Then, $i \longleftrightarrow i'$ in $\mathcal{G}^\infty$. The proof is completed.

Proof of Lemma 3. We prove a more general version of the first statement in Lemma 3 with the dimension of β being $K \geq 1$. Define $\delta(N) = \left(\frac{\ln N}{cC_k N}\right)^{\frac{1}{K+1}}$ where $C_k = \frac{\pi^{(K+1)/2}}{\Gamma((K+3)/2)}$ and c is the constant in Assumption 4.

First, given any (α, β), we examine the probability of the event $\mathcal{E}(\alpha, \beta, N) := \{\min_{1 \leq i \leq N} \|(\alpha, \beta) - (\alpha_{i0}, \beta_{i0})\| > \delta(N)\}$. Under Assumption 4,

$$\Pr\left(\mathcal{E}(\alpha, \beta, N)\right) = \prod_{i=1}^{N} \Pr\left(\|(\alpha, \beta) - (\alpha_{i0}, \beta_{i0})\| > \delta(N)\right) \leq \left(1 - \frac{\ln N}{N}\right)^N.$$

Second, because $\Omega_{\alpha,\beta}$ is compact (say, the unit square), for a given N, we can always find up to $R(\Omega_{\alpha,\beta})N/(\ln N)$ sets in $\{\text{Ball}_{(\alpha,\beta)}(\delta(N))\}_{(\alpha,\beta)\in\Omega_{\alpha,\beta}}$ such that their union covers $\Omega_{\alpha,\beta}$ where $R(\Omega_{\alpha,\beta})$ is a constant that only depends on $\Omega_{\alpha,\beta}$ and c. Denote by $\mathcal{B}_s$ each of the finite sets. Now consider the probability of the event

$$\begin{aligned} &\{\sup_{(\alpha,\beta)\in\Omega_{\alpha,\beta}} \min_{1\leq i\leq N} \|(\alpha, \beta) - (\alpha_{i0}, \beta_{i0})\| > 2\delta(N)\} \\ &\quad \subset \cup_{s=1}^{\frac{R(\Omega_{\alpha,\beta})N}{(\ln N)}} \{\sup_{(\alpha,\beta)\in\mathcal{B}_s} \min_{1\leq i\leq N} \|(\alpha, \beta) - (\alpha_{i0}, \beta_{i0})\| > 2\delta(N)\}. \end{aligned}$$

The probability of $\{\sup_{(\alpha,\beta)\in\mathcal{B}_s}\min_{1\leq i\leq N}\|(\alpha,\beta)-(\alpha_{i0},\beta_{i0})\|>2\delta(N)\}$ is bounded by $(1-\ln N/N)^N$ that corresponds to the probability of none of the (α_{i0},β_{i0}) falling in $\mathcal{B}_s$. Then,

$$\begin{aligned}&\Pr\left(\sup_{(\alpha,\beta)\in\Omega_{\alpha,\beta}}\min_{1\leq i\leq N}\|(\alpha,\beta)-(\alpha_{i0},\beta_{i0})\|>2\delta(N)\right)\\&\leq\frac{R(\Omega_{\alpha,\beta})N}{\ln N}\left(1-\frac{\ln N}{N}\right)^N\sim\frac{R(\Omega_{\alpha,\beta})}{\ln N}\to 0.\end{aligned}$$

The proof of the second statement differs from that of the first one in that ξ_{t0} is not independent across t, but only admits a weak dependence specified in Assumption 4. Define $\delta(T)=\frac{1}{2cT^{\frac{1}{3}}}$. Note that for any t and $\xi\in\Omega_\xi$, $\Pr(|\xi-\xi_{t0}|\geq\delta(T))=1-\Pr(|\xi-\xi_{t0}|<\delta(T))\leq 1-2c\delta(T)$. To adapt the proof of the first statement:

$$\begin{aligned}\Pr\left(\min_{1\leq t\leq T}|\xi-\xi_{t0}|\geq\delta(T)\right)&\leq\Pr\left(|\xi-\xi_{t0}|\geq\delta(T),\ t=1,1+T^{\frac{1}{3}(1/\mu+1)},...,1+\lfloor\frac{T-1}{T^{\frac{1}{3}(1/\mu+1)}}\rfloor T^{\frac{1}{3}(1/\mu+1)}\right)\\&\leq\Pr\left(|\xi-\xi_{t0}|\geq\delta(T),\ t=1+T^{\frac{1}{3}(1/\mu+1)},...,1+\lfloor\frac{T-1}{T^{\frac{1}{3}(1/\mu+1)}}\rfloor T^{\frac{1}{3}(1/\mu+1)}\right)\Pr\left(|\xi-\xi_{10}|\geq\delta(T)\right)+a(T^{\frac{1}{3}(1/\mu+1)})\\&\leq\Pr\left(|\xi-\xi_{t0}|\geq\delta(T),\ t=1+2T^{\frac{1}{3}(1/\mu+1)},...,1+\lfloor\frac{T-1}{T^{\frac{1}{3}(1/\mu+1)}}\rfloor T^{\frac{1}{3}(1/\mu+1)}\right)\Pr\left(|\xi-\xi_{1+T^{\frac{1}{3}(1/\mu+1)}0}|\geq\delta(T)\right)\Pr\left(|\xi-\xi_{10}|\geq\delta(T)\right)\\&+a(T^{\frac{1}{3}(1/\mu+1)})\left(1+\Pr\left(|\xi-\xi_{10}|\geq\delta(T)\right)\right)\\&\leq\Pr\left(|\xi-\xi_{10}|\geq\delta(T)\right)\Pr\left(|\xi-\xi_{1+T^{\frac{1}{3}(1/\mu+1)}0}|\geq\delta(T)\right)...\Pr\left(|\xi-\xi_{1+\lfloor\frac{T-1}{T^{\frac{1}{3}(1/\mu+1)}}\rfloor T^{\frac{1}{3}(1/\mu+1)}0}|\geq\delta(T)\right)+\frac{a(T^{\frac{1}{3}(1/\mu+1)})}{2c\delta(T)}\\&\leq(1-2c\delta(T))^{\lfloor\frac{T-1}{T^{\frac{1}{3}(1/\mu+1)}}\rfloor}+\frac{a(T^{\frac{1}{3}(1/\mu+1)})}{2c\delta(T)}.\end{aligned}$$

Following the same strategy as the proof of the first statement, we obtain that

$$\begin{aligned}&\Pr\left(\sup_{\xi\in\Omega_\xi}\min_{1\leq t\leq T}|\xi-\xi_{t0}|>2\delta(T)\right)\quad\leq\frac{R(\Omega_\xi)}{\delta(T)}\left((1-2c\delta(T))^{\lfloor\frac{T-1}{T^{\frac{1}{3}(1/\mu+1)}}\rfloor}+\frac{a(T^{\frac{1}{3}(1/\mu+1)})}{2c\delta(T)}\right)\\&=2cR(\Omega_\xi)\left(\underbrace{\frac{(1-2c\delta(T))^{\lfloor\frac{T-1}{T^{\frac{1}{3}(1/\mu+1)}}\rfloor}}{2c\delta(T)}}_{\sim T^{1/3}\exp\{-T^{\frac{1}{3}(1-1/\mu)}\}}+\underbrace{\frac{a(T^{\frac{1}{3}(1/\mu+1)})}{(2c\delta(T))^2}}_{\leq O(T^{\frac{1}{3}(1-\mu)})}\right)\\&\to 0.\end{aligned}$$

The proof is completed.

G Monte Carlo: Additional Results

Results on coefficients estimates. In the MLE with polynomial sieves in Section 4, we normalize $(\beta^{(1)},\alpha_1^{(1)},\alpha_1^{(2)})$ by their true values. As discussed in footnote

Table 4: Finite-sample performances of the sieve MLE when $g_0(\delta) = \frac{\exp\{2\exp\{\delta\}\}}{1+\exp\{2\exp\{\delta\}\}}$

Polynomial sieve	$\beta_i^{(2)}$	$\alpha_i^{(1)}$	$\alpha_i^{(2)}$
$N = 50$, $d = 1$	1.2807	0.2750	0.4599
$d = 2$	1.6370	0.5189	0.8434
$d = 3$	1.6271	0.5074	0.7889
$d = 4$	1.6186	0.5074	0.7750
$N = 100$, $d = 1$	0.8724	0.1918	0.2835
$d = 2$	0.8593	0.2807	0.6784
$d = 3$	0.8104	0.1796	0.4755
$d = 4$	0.8103	0.1721	0.4627
$N = 200$, $d = 1$	0.5692	0.1103	0.1916
$d = 2$	0.5509	0.4599	0.8361
$d = 3$	0.4147	0.0547	0.1269
$d = 4$	0.4235	0.0586	0.1326

Notes: Each cell corresponds to the average distance metrics between the estimated object and its true value over 200 repetitions for a given sample size N and MLE with polynomial sieves of degree $d = 1, ..., 4$.

8, together with these normalizations, Theorem 3 implies that the estimators for $(\beta_i^{(2)}, \alpha_i^{(1)}, \alpha_i^{(2)})$ converge to their true values as N increases to infinity. In Table 4, we report their distance metrics defined as $\sqrt{\frac{\sum_{i=1}^N (\hat{\beta}_i^{(2)} - \beta_{i0}^{(2)})^2}{N}}$, $\sqrt{\frac{\sum_{i=1}^N (\hat{\alpha}_i^{(1)} - \alpha_{i0}^{(1)})^2}{N}}$, and $\sqrt{\frac{\sum_{i=1}^N (\hat{\alpha}_i^{(2)} - \alpha_{i0}^{(2)})^2}{N}}$, for the scenario $g_0(\delta) = \frac{\exp\{2\exp\{\delta\}\}}{1+\exp\{2\exp\{\delta\}\}}$. For each sieve dimension d and each object of interest, the corresponding distance metric decreases as N increase from 50 to 200.

Results on alternative data generating process. We consider the same static binary choice model as that in section 4 with a different generating process for covariates: for $1 \leq i < j \leq N$,

$$\Pr(y_{ij} = 1 | w_{ij}, \beta_i, \alpha_i^{(1)}, \alpha_j^{(2)}) = g_0(w_{ij}^{(1)}\beta^{(1)} + w_{ij}^{(2)}\beta_i^{(2)} + \alpha_i^{(1)} + \alpha_j^{(2)}),$$

where $w_{ij}^{(1)} = |z_i - z_j|$ and $w_{ij}^{(2)} = |x_i - x_j|$. x_i and z_i are respectively i.i.d. $\mathcal{U}[-1, 1]$ and $\mathcal{U}[-1, 0]$ across $i = 1, \ldots, N$. Individual-specific slopes $\beta_i^{(2)} = -0.2 - 0.1x_i - \mu_i$ where μ_i are i.i.d. $\mathcal{U}[0, 1]$. Fixed effects $\alpha_i^{(1)} = 0.5x_i + \eta_i^{(1)}$ and $\alpha_i^{(2)} = 0.5x_i + \eta_i^{(2)}$ where $\eta_i^{(1)}$ ($\eta_i^{(2)}$) are i.i.d. $\mathcal{U}[-0.5, 0.5]$. We set $\beta^{(1)} = 1$. $(x_i)_{i=1}^N$, $(z_i)_{i=1}^N$, $(\mu_i)_{i=1}^N$, $(\eta_i^{(1)})_{i=1}^N$, and $(\eta_i^{(2)})_{i=1}^N$ are independent. Note that $(w_{ij}^{(1)}, w_{ij}^{(2)})_{j \neq i}$

Table 5: Finite-sample performances: Polynomial sieves

Scenario $g_0(\delta) =$	$\frac{\exp\{\delta\}}{1+\exp\{\delta\}}$			$\Phi(\delta)$			$\frac{\exp\{2\exp\{\delta\}\}}{1+\exp\{2\exp\{\delta\}\}}$		
	$\text{AME}_i^{(1)}$	$\text{AME}_i^{(2)}$	g	$\text{AME}_i^{(1)}$	$\text{AME}_i^{(2)}$	g	$\text{AME}_i^{(1)}$	$\text{AME}_i^{(2)}$	g
$N = 50$, Logit	0.0426	0.1729	×	0.0493	0.1554	×	0.0720	0.1069	×
Probit	0.0417	0.1804	×	0.0491	0.1676	×	0.0633	0.1223	×
Poly. sieve, $d = 1$	0.0363	0.1744	0.0947	0.0504	0.1678	0.0974	0.0729	0.1321	0.1485
$d = 2$	0.0575	0.1776	0.1042	0.0560	0.1689	0.0991	0.0560	0.1122	0.0992
$d = 3$	0.0729	0.1880	0.1110	0.0742	0.1913	0.1157	0.0586	0.1134	0.0995
$d = 4$	0.0853	0.1968	0.1232	0.0830	0.1975	0.1194	0.0596	0.1162	0.0949
$N = 100$, Logit	0.0240	0.1188	×	0.0309	0.1075	×	0.0559	0.0859	×
Probit	0.0229	0.1196	×	0.0290	0.1098	×	0.0479	0.0911	×
Poly. sieve, $d = 1$	0.0233	0.1188	0.0700	0.0314	0.1102	0.0658	0.0537	0.0991	0.1245
$d = 2$	0.0264	0.1194	0.0737	0.0323	0.1104	0.0665	0.0383	0.0779	0.0961
$d = 3$	0.036804	0.1272	0.0791	0.0306	0.1106	0.0637	0.0371	0.0774	0.0945
$d = 4$	0.0528	0.1353	0.1031	0.0319	0.1112	0.0634	0.0369	0.0777	0.0928
$N = 200$, Logit	0.0144	0.0784	×	0.0209	0.0723	×	0.0482	0.0714	×
Probit	0.0137	0.0785	×	0.0186	0.0721	×	0.0431	0.0716	×
Poly. sieve, $d = 1$	0.0144	0.0784	0.0413	0.0210	0.0725	0.0358	0.0479	0.0772	0.1145
$d = 2$	0.0151	0.0785	0.0426	0.0210	0.0724	0.0364	0.0278	0.0566	0.1126
$d = 3$	0.0158	0.0786	0.0431	0.0188	0.0722	0.0358	0.0230	0.0562	0.0817
$d = 4$	0.0166	0.0788	0.0473	0.0191	0.0723	0.0353	0.0233	0.0562	0.0780

Notes: Each cell corresponds to the average distance metrics between the estimated object and its true value over 200 repetitions for a given sample size N, scenario of true link function g_0, and the model used in the MLE (logit, probit, or polynomial sieves of degree $d = 1, \ldots, 4$). For $\text{AME}_i^{(k)}$ with $k = 1, 2$, the distance metrics is defined as $\sqrt{\sum_{i=1}^N (\widehat{\text{AME}}_i^{(k)} - \text{AME}_{i0}^{(k)})^2/N}$ where 0 refers to the true values. For the sieve MLE, the distance corresponding to the link function is defined as $\sqrt{\sum_{m=1}^M \left(\hat{g}(\delta_m) - g_0(\delta_m)\right)^2/M}$ where $(\delta_m)_{m=1}^M$ is an equal-spaced (by 0.1) sequence of values covering the true range of the index in the data generating process.

are not independent across i's and Assumption 1(c) is violated. The results are summarized in Table 5. Despite the violation, the sieve MLE seems to have an analogous performance to the one under the data generating process in section 4. One potential reason is that the identification and consistency results could hold under weaker (or alternative) assumptions than those proposed in the paper. We leave this possibility for future research.

H Acknowledgment

We are grateful to Mingli Chen, Xavier D'Haultfœuille, Christophe Gaillac, Ivana Komunjer, Eric Renault, Amrei Stammann, Francis Vella, and Martin Weidner for their useful suggestions. We also thank the participants at the 2021 Bristol Econometric Study Group, 2021 European Winter Meeting of the Econometric Society (ES), Encounters in Econometric Theory at Oxford, 2022 ES Asia Meeting, CREST, CUFE, Georgetown, LSE, Oxford, University of Chicago, and Warwick for their helpful comments. Martin Mugnier gratefully acknowledges financial support from the research grants Otelo (ANR-17-CE26-0015-041) and ANR "In-

vestissements d'avenir" (ANR-18-EURE-0005, ANR-11-LABX-0047, and ANR-17-EURE-0001).

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67(2), 251–333.

Aghion, P., J. Van Reenen, and L. Zingales (2013): "Innovation and institutional ownership," *American economic review*, 103(1), 277–304.

Andrews, D. W. K. (1984): "Non-Strong Mixing Autoregressive Processes," *Journal of Applied Probability*, 21(4), 930–934.

Arellano, M., and S. Bonhomme (2011): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *The Review of Economic Studies*, 79(3), 987–1020.

Bai, J. (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77(4), 1229–1279.

Berry, S., A. Gandhi, and P. Haile (2013): "Connected substitutes and invertibility of demand," *Econometrica*, 81(5), 2087–2111.

Boneva, L., and O. Linton (2017): "A discrete-choice model for large heterogeneous panels with interactive fixed effects with an application to the determinants of corporate bond issuance," *Journal of Applied Econometrics*, 32(7), 1226–1243.

Candelaria, L. E. (2020): "A Semiparametric Network Formation Model with Unobserved Linear Heterogeneity," .

Chamberlain, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18(1), 5–46.

Charalambos, D., and B. Aliprantis (2013): *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag Berlin and Heidelberg GmbH & Company KG.

Charbonneau, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20(3), S1–S13.

Chen, M. (2016): "Estimation of nonlinear panel models with multiple unobserved effects," Discussion paper.

Chen, M., I. Fernández-Val, and M. Weidner (2021): "Nonlinear factor models for network and panel data," *Journal of Econometrics*, 220(2), 296–324.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2009): "Improving point and interval estimators of monotone functions by rearrangement," *Biometrika*, 96(3), 559–575.

de Paula, A. (2020): "Econometric Models of Network Formation," *Annual Review of Economics*, 12(1), 775–799.

Dhaene, G., and K. Jochmans (2015): "Split-panel jackknife estimation of fixed-effect models," *The Review of Economic Studies*, 82(3), 991–1030.

D'Haultfoeuille, X., S. Hoderlein, and Y. Sasaki (2021): "Testing and relaxing the exclusion restriction in the control function approach," *Journal of Econometrics*.

D'Haultfœuille, X., A. Wang, P. Février, and L. Wilner (2022): "Estimating the Gains (and Losses) of Revenue Management," *arXiv preprint arXiv:2206.04424*.

Dubois, P., R. Griffith, and M. O'Connell (2020): "How well targeted are soda taxes?," *American Economic Review*, 110(11), 3661–3704.

Dudley, R. M. (2018): *Real analysis and probability*. CRC Press.

Fan, J., and Q. Yao (2003): *Nonlinear time series: nonparametric and parametric methods*, vol. 20. Springer.

Fernández-Val, I. (2009): "Fixed effects estimation of structural parameters and marginal effects in panel probit models," *Journal of Econometrics*, 150(1), 71–85.

Fernández-Val, I., and J. Lee (2013): "Panel data models with nonadditive unobserved heterogeneity: Estimation and inference," *Quantitative Economics*, 4(3), 453–481.

Fernández-Val, I., and M. Weidner (2016): "Individual and time effects

in nonlinear panel models with large N, T," *Journal of Econometrics*, 192(1), 291–312.

——— (2018): "Fixed Effects Estimation of Large-T Panel Data Models," *Annual Review of Economics*, 10(1), 109–138.

Freyberger, J., and M. A. Masten (2019): "A practical guide to compact infinite dimensional parameter spaces," *Econometric Reviews*, 38(9), 979–1006.

Gale, D., and H. Nikaido (1965): "The Jacobian matrix and global univalence of mappings," *Mathematische Annalen*, 159(2), 81–93.

Gallant, A. R., and D. W. Nychka (1987): "Semi-nonparametric maximum likelihood estimation," *Econometrica: Journal of the econometric society*, pp. 363–390.

Gao, J., F. Liu, B. Peng, and Y. Yan (2023): "Binary response models for heterogeneous panel data with interactive fixed effects," *Journal of Econometrics*, 235(2), 1654–1679.

Gao, W. Y. (2020): "Nonparametric identification in index models of link formation," *Journal of Econometrics*, 215(2), 399–413.

Graham, B. S. (2017): "An Econometric Model of Network Formation With Degree Heterogeneity," *Econometrica*, 85(4), 1033–1063.

Graham, B. S., and J. L. Powell (2012): "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models," *Econometrica*, 80(5), 2105–2152.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both n and T Are Large," *Econometrica*, 70(4), 1639–1657.

Hahn, J., and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Hansen, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24(3), 726–748.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating trade flows: Trading partners and trading volumes," *The Quarterly Journal of Economics*, 123(2), 441–487.

Hsiao, C., and M. Pesaran (2004): "Random Coefficient Panel Data Models," Cambridge Working Papers in Economics 0434, Faculty of Economics, University of Cambridge.

Jochmans, K. (2018): "Semiparametric analysis of network formation," *Journal of Business & Economic Statistics*, 36(4), 705–713.

Jochmans, K., and M. Weidner (2019): "Fixed-Effect Regressions on Network Data," *Econometrica*, 87(5), 1543–1560.

——— (2024): "Inference on a distribution from noisy draws," *Econometric Theory*, 40(1), 60–97.

Lancaster, T. (2000): "The incidental parameter problem since 1948," *Journal of econometrics*, 95(2), 391–413.

Latała, R. (2005): "Some estimates of norms of random matrices," *Proceedings of the American Mathematical Society*, 133(5), 1273–1282.

Lei, L., and B. Ross (2024): "Estimating Counterfactual Matrix Means with Short Panel Data," .

Lewbel, A. (2014): "An overview of the special regressor method," .

Newey, W. K. (1991): "Uniform convergence in probability and stochastic equicontinuity," *Econometrica: Journal of the Econometric Society*, pp. 1161–1167.

Neyman, J., and E. L. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica: Journal of the Econometric Society*, pp. 1–32.

Rio, E. (1993): "Covariance inequalities for strongly mixing processes," in *Annales de l'IHP Probabilités et statistiques*, vol. 29, pp. 587–597.

Swamy, P. A. V. B. (1970): "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38(2), 311–323.

Toth, P. (2017): "Semiparametric estimation in network formation models with homophily and degree heterogeneity," *Available at SSRN 2988698*.

Vytlacil, E., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75(3), 757–779.

Williams, B. (2020): "Nonparametric identification of discrete choice models

with lagged dependent variables," *Journal of Econometrics*, 215(1), 286–304.

Zeleneev, A. (2020): "Identification and Estimation of Network Models with Nonparametric Unobserved Heterogeneity," Discussion paper.