Econometric Theory, **0**, 2024, 1–40.
doi:10.1017/S0266466624000057

SEMIPARAMETRIC ESTIMATION OF DYNAMIC BINARY CHOICE PANEL DATA MODELS

FU OUYANG
University of Queensland

THOMAS TAO YANG
Australian National University

We propose a new approach to the semiparametric analysis of panel data binary choice models with fixed effects and dynamics (lagged dependent variables). The model under consideration has the same random utility framework as in Honoré and Kyriazidou (2000, *Econometrica* 68, 839–874). We demonstrate that, with additional serial dependence conditions on the process of deterministic utility and tail restrictions on the error distribution, the (point) identification of the model can proceed in two steps, and requires matching only the value of an index function of explanatory variables over time, rather than the value of each explanatory variable. Our identification method motivates an easily implementable, two-step maximum score (2SMS) procedure – producing estimators whose rates of convergence, in contrast to Honoré and Kyriazidou's (2000, *Econometrica* 68, 839–874) methods, are independent of the model dimension. We then analyze the asymptotic properties of the 2SMS procedure and propose bootstrap-based distributional approximations for inference. Evidence from Monte Carlo simulations indicates that our procedure performs satisfactorily in finite samples.

1. INTRODUCTION

In this paper, we propose a two-step estimation method for panel data binary choice models with fixed effects and dynamics. Specifically, we consider binary choice models of the form

$$y_{it} = 1\left[x_{it}'\beta + \gamma y_{it-1} + \alpha_i - \epsilon_{it} > 0\right], i = 1, \ldots, n, t = 1, \ldots, T, \text{[1]} \quad \textbf{(1.1)}$$

We thank the editor (Peter C. B. Phillips), the co-editor (Iván Fernández-Val), and the two anonymous referees for their valuable comments and suggestions, which have significantly improved the quality of this paper. We are grateful to Yonghong An, Shakeeb Khan, Arthur Lewbel, Takuya Ura, Hanghui Zhang, and Yichong Zhang for their insightful feedback and discussions. We also thank participants at the 2019 Asian Meeting of the Econometric Society, the 2019 Shanghai Workshop of Econometrics, and the 2022 Australasia Meeting of the Econometric Society for their helpful comments. Fu Ouyang acknowledges the financial support provided by the Faculty of Business, Economics, and Law (BEL) at the University of Queensland through the 2019 BEL New Staff Research Start-Up Grant. All errors are our responsibility. Address correspondence to Fu Ouyang, School of Economics, University of Queensland, St Lucia, QLD, Australia, e-mail: f.ouyang@uq.edu.au

[1] The identification approach and estimation method presented in this paper can be applied to models with unbalanced panels as long as the unbalancedness is not the result of endogenous attrition.

where T is small and n is large, x_{it} is a $K \times 1$ vector of (time-varying) explanatory variables,[2] y_{it-1} is the lagged dependent variable, α_i represents a time-invariant, individual-specific (fixed) effect, and ϵ_{it} is an idiosyncratic error term. Both α_i and ϵ_{it} are unobservable to the econometrician. Following Honoré and Kyriazidou (2000) (referred to as HK henceforth), we assume the strong exogeneity that $(x_{i1}, \ldots, x_{iT}) \perp (\epsilon_{i1}, \ldots, \epsilon_{iT}) | \alpha_i$ and that ϵ_{it} are independent and identically distributed (i.i.d.) across t conditional on α_i. Interest centers on estimating the preference parameter $\theta \equiv (\beta', \gamma)'$. y_{i0} is assumed to be observed, although the model is not specified in the initial period 0. In the literature, the lagged terms y_{it-1} and the fixed effect α_i are referred to as the "state dependence" (see Heckman, 1981a, 1981b) and the "unobservable heterogeneity,"respectively. The co-existence of these two terms complicates the identification and estimation of θ, owing to the multiple sources of persistence in y_{it}.

This paper resembles other panel data discrete response literature using fixed effects methods, in that we impose no restrictions on the distribution of α_i, conditional on the observed explanatory variables. Arellano and Honoré (2001) review early works on estimating β in model (1.1) with no state dependence (y_{it-1}). Chamberlain (2010) shows that, outside of the logistic case, these static binary choice models have a zero information bound, and the identification requires that at least one of the observed covariates have unbounded support. In the presence of lagged dependent variables, various conditional maximum likelihood methods have been developed for variants of model (1.1) with logistic errors and at least four observations ($T \geq 3$) per individual.[3] Honoré and De Paula (2021) provide a comprehensive review of this literature.[4] Several new methods have been proposed for dynamic Logit models based on moment conditions. Leading examples include Honoré and Weidner (2020), Dobronyi, Gu, and Kim (2021), Kitazawa (2022), and Dano (2023), among others.

HK is the first to consider the semiparametric identification and estimation of model (1.1). They demonstrate that θ can be identified if, in addition to assumptions analogous to those in Manski (1987), all explanatory variables are strictly exogenous, ϵ_{it}'s are serially independent, and $T \geq 3$. However, the rate of convergence of their estimator decreases as the number of continuous regressors increases, and is slower than the standard maximum score (MS) rate derived by Kim and Pollard (1990).

There are several alternative fixed effects approaches to the semi- and nonparametric analysis of dynamic binary choice models. Honoré and Lewbel (2002) propose an identification strategy that requires an exclusion restriction (excluded regressor). Chen, Khan, and Tang (2019) show that the exclusion restriction in Honoré and Lewbel (2002) implicitly assumes (conditional) serial independence

[2] Any time-invariant covariates can be thought of as being part of the fixed effect α_i.

[3] Throughout this paper, this means the data contain y_{i0} and $(y_{i1}, y_{i2}, y_{i3}, x_{i1}, x_{i2}, x_{i3})$ for each individual i.

[4] Works such as Bartolucci and Nigro (2010, 2012) and Al-Sadoon, Li, and Pesaran (2017) study the estimation of model (1.1) under alternative specifications.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

of the excluded regressor. Williams (2019) studies the nonparametric identification of dynamic binary choice models that satisfy certain exclusion restrictions. In the absence of excluded regressors, some recent works, such as Khan, Ponomareva, and Tamer (2020) and Aristodemou (2021), characterize the (sharp) identified set for θ under mild conditions. We refer interested readers to the survey article by Honoré and De Paula (2021) and Chapter 7 in Hsiao (2022) for a detailed review of this literature.

This paper takes one step in the direction of HK's semiparametric estimator, in the sense that we provide sufficient conditions under which model (1.1) can be identified and estimated, without needing to match each of the explanatory variables over time, provided we have at least five observations per individual are observed (i.e., $T \geq 4$).[5] The key insight here is that the identification of θ can proceed in two steps. First, β can be identified based on sequences of $\{y_{it}\}$, for which $y_{is-1} = y_{it-1}$ and $y_{is+1} = y_{it+1}$ for some $1 \leq s < t \leq T-1$ with $t \geq s+2$ (e.g., in the simplest case where $T = 4$, β is identified based on observations with $y_0 = y_2 = y_4$), if the distribution of explanatory variables x_{it} satisfies certain serial dependence and stochastic dominance restrictions. Then, using the identified β, γ can be identified by simply matching $x_{it}'\beta$ over time. We propose an estimation procedure for β and γ, establish the asymptotics for our estimators, and provide an inference method that uses the bootstrap. We investigate their finite-sample properties using Monte Carlo experiments.

As demonstrated by Honoré and Tamer (2006), matching exogenous utilities over time seems to be essential for the point identification in "distribution-free" dynamic discrete choice models.[6] However, the approach proposed here involves matching an identified linear combination of x_{it}, rather than HK's matching each component of x_{it}. Consequently, in contrast to the results presented by HK, the rates of convergence of our proposed estimators are independent of the dimension of the regressor space, making our approach particularly useful for models with a higher-dimensional design.

It is known that panel data binary choice models with unobserved heterogeneity and dynamics can be estimated using the random effects or correlated random coefficients approach. Examples include Arellano and Carrasco (2003), Wooldridge (2005), and Honoré and Tamer (2006). In addition to preference parameters, these approaches often allow the econometrician to calculate other quantities of interest, such as choice probabilities and marginal effects. However, these approaches require the specification of the statistical relation between the explanatory variables and α_i. Further, they require one to specify the distribution of y_{i0}, conditional on the observed explanatory variables and α_i, which raises the so-called initial condition problem. Conversely, the fixed effects approaches attempt

[5] That is, at least y_{i0} and $(y_{i1}, y_{i2}, y_{i3}, y_{i4}, x_{i1}, x_{i2}, x_{i3}, x_{i4})$ are observed for each individual i. This is a restriction on the minimum panel length, which is satisfied for many longitudinal panel data sets.

[6] More precisely, Honoré and Tamer (2006) provided examples of point identification failure when it is impossible to match x_{it} over time.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

to estimate preference parameters without making these subtle specifications. Finally, there is also literature exploring the identification and estimation of various partial effects in panel data models (see, e.g., Altonji and Matzkin, 2005; Chernozhukov et al., 2013, and more recent advancements by Torgovitsky, 2019; Aguirregabiria and Carro, 2021; Dobronyi et al., 2021; Davezies, D'Haultfoeuille, and Laage, 2022; Liu, Poirier, and Shiu, 2023, among others).

Dynamic binary choice models have a wide range of applications, including the study of labor force participation (Damrongplasit, Hsiao, and Zhao, 2018), poverty dynamics (Biewen, 2009)), health status (Halliday, 2008), educational attainment (Cameron and Heckman, 1998, 2001), stock market participation (Alessie, Hochguertel, and Soest, 2004), brand loyalty (Chintagunta, Kyriazidou, and Perktold, 2001), welfare participation (Chay, Hoynes, and Hyslop, 1999), and firm behavior (Kerr, Lincoln, and Mishra, 2014), among others. Most applications typically employ parametric forms of the model (1.1), such as Logit and Probit, or random effects assumptions. The robustness from the distribution-free and fixed effects specification makes the approach proposed here a competitive alternative to existing parametric and random effects methods. Note that the theoretical validity of our approach relies on certain restrictions on the serial dependence of explanatory variables. Thus, before applying our method, we suggest that applied researchers detrend and seasonally adjust x_{it} in a way that makes them resemble "white noise" conditional on α_i.[7]

The remainder of this paper is organized as follows: Section 2 establishes the identification of θ under different sets of sufficient conditions, based on which, a two-step maximum score (2SMS) procedure is proposed in Section 3. Sections 4 and 5 derive the asymptotic properties of the 2SMS estimator and propose bootstrap-based inference methods, respectively. We present the results of Monte Carlo experiments in Section 6 that examine the finite-sample performance of our proposed method. Section 7 concludes the paper. We prove the main theorems and present the main simulation results in the Appendixes. The Supplementary Material to this paper includes proofs of all technical lemmas, technical details for the bootstrap inference, and results for supplementary simulation studies.

For ease of reference, we next describe the notation maintained throughout this paper.

Notation. All vectors are column vectors. $\mathbb{R}^p$ is a p-dimensional Euclidean space equipped with the Euclidean norm $\|\cdot\|_2$. We reserve the letter $i \in \mathcal{N} \equiv \{1,\ldots,n\}$ for indexing individuals, and the letters $s,t \in \mathcal{T} \equiv \{1,\ldots,T\}$ for indexing time periods. An observation is indexed by (i,t). Vector x_{its} denotes $x_{it} - x_{is}$. The first element of x_{its} is denoted by $x_{its,1}$ and the sub-vector comprising its remaining elements is denoted by $\tilde{x}_{its}$. As is common in the panel data literature, we use

[7]Monte Carlo results show that relaxing these restrictions does not significantly affect our estimators' finite-sample performances (see Section 6 for a more detailed discussion).

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

the notation ξ^t to denote $\left(\xi_1', \ldots, \xi_t'\right)'$. For example, suppressing the subscript i, $y^t \equiv (y_1, \ldots, y_t)'$, a $t \times 1$ vector. $F_{\zeta|\cdot}$ and $f_{\zeta|\cdot}$ denote, respectively, the conditional cumulative distribution function (CDF) and probability density function (PDF) of a random vector ζ conditional on $\cdot$. For two random vectors, u and v, the notation $u \overset{d}{=} v|\cdot$ means that u and v have identical distributions, conditional on $\cdot$, and $u \perp v|\cdot$ means that u and v are independent, conditional on $\cdot$. We use $P(\cdot)$ and $\mathbb{E}[\cdot]$ to denote probability and expectation, respectively. Function $1[\cdot]$ is an indicator function, equal to one when the event in the brackets is true, and zero otherwise. Function $sgn(\cdot)$ denotes the sign function, equal to 1 when $\cdot$ is positive, 0 when $\cdot$ is 0, and -1 when $\cdot$ is negative. Symbols $\setminus$, $'$, $\propto$, $\Leftrightarrow$, $\overset{d}{\to}$, and $\overset{P}{\to}$ represent set difference, matrix transposition, proportionality, "if and only if," convergence in distribution, and convergence in probability, respectively. For any (random) positive sequences, $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ ($O_P(b_n)$) means that a_n/b_n is bounded (bounded in probability), and $a_n = o(b_n)$ ($o_P(b_n)$) means that $a_n/b_n \to 0$ ($a_n/b_n \overset{P}{\to} 0$).

2. IDENTIFICATION

This section provides sufficient conditions for identifying the parameter θ with no need to match observed covariates x_{it} over time. Under these assumptions, we derive a set of identification inequalities that can be taken to data for (point) estimation and inference on the parameter θ.

To simplify the notation, we suppress the subscript i in the rest of this paper whenever it is clear from the context that all variables relate to each individual. Suppose that a random sample from a population of independent individuals[8] is observed for $T+1$ ($= |\mathcal{T} \cup \{0\}|$) periods. Recall that, for all $t \in \mathcal{T}$,

$$y_t = 1\left[x_t'\beta + \gamma y_{t-1} + \alpha - \epsilon_t > 0\right]. \tag{2.1}$$

Note that the model is incomplete, in the sense that it does not specify the relationship between y_0 and $(x^T, \alpha, \epsilon^T)$. This is known as the initial condition problem in panel data literature. This paper uses a fixed effects approach, in which we attempt to estimate $\theta = (\beta', \gamma)'$ without making any assumptions on the distribution of α, conditional on explanatory variables. This helps us to avoid explicitly specifying the functional form of $p_0(x^T, \alpha) \equiv P(y_0 = 1|x^T, \alpha)$, and thus circumvents the initial condition problem.

As mentioned, we impose no restriction on $F_{\alpha|x^T}$, but place the following restrictions on observed covariates x^T and unobserved idiosyncratic errors ϵ^T.

Assumption A. For all α and $s, t \in \mathcal{T}$,

(a) (i) $\epsilon^T \perp (x^T, y_0)|\alpha$, (ii) $\epsilon_s \perp \epsilon_t|\alpha$, and (iii) $\epsilon_s \overset{d}{=} \epsilon_t|\alpha$.
(b) $F_{\epsilon_t|\alpha}$ is absolutely continuous with PDF $f_{\epsilon_t|\alpha}$ and support $\mathbb{R}$.

[8] Here, the term "independent individuals" refers to the assumption that $(\alpha_i, y_{i0}, x_{i1}, \ldots, x_{iT}, \epsilon_{i1}, \ldots, \epsilon_{iT})$ is independently distributed across i.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

(c) (i) One of the regressors, without loss of generality (w.l.o.g.) $x_{ts,1}$, has almost everywhere positive probability density on $\mathbb{R}$, conditional on $\tilde{x}_{ts}$ and α, and (ii) the coefficient β_1 on $x_{ts,1}$ is nonzero.
(d) The support $\mathcal{X}_{ts}$ of $F_{x_{ts}|\alpha}$ is not contained in any proper linear subspace of $\mathbb{R}^K$.
(e) $\theta = (\beta', \gamma)' \in \mathcal{B} \times \text{int}(\mathcal{R})$, where $\mathcal{B} \equiv \{b = (b_1, \ldots, b_K)' \in \mathbb{R}^K | \|b\|_2 = 1\}$ and $\mathcal{R}$ is a compact subset of $\mathbb{R}$ with a non-empty interior.

Assumption A places the same set of restrictions on the joint distribution of $(x^T, \alpha, \epsilon^T)$ as in HK. While not stated explicitly in their Theorem 4, HK use Assumption A(a), the exogeneity of (x^T, y_0) and serial independence of $\{\epsilon_t\}$, conditional on α, to derive the moment inequalities for the identification. Note that Assumption A(a) implies that the fixed effects α pick up two types of dependence in the model: the dependence over time in the unobservables, and the dependence between explanatory variables and unobservables. As a result, in model (2.1), ϵ_t is independent of $\left(x^T, y^{t-1}\right)$, conditional on α. Furthermore, Assumption A(a) is a special case of the group homogeneity restriction, $\epsilon_s \overset{d}{=} \epsilon_t | (x_s, x_t, \alpha)$, imposed in Manski (1987), Pakes and Porter (2016), and Shi, Shum, and Song (2018) for identifying static discrete choice models (without controlling the lagged term y_{t-1} in the model). Thus, we can suppress the time subscript t in $F_{\epsilon_t|\alpha}$ and $f_{\epsilon_t|\alpha}$ in the rest of this paper without ambiguity. Assumption A(b) is a regularity condition that ensures that both $y_s \neq y_t$ and $y_s = y_t$ occur with positive probabilities for all α and $s, t \in \mathcal{T}$.

It is known and documented in the relevant literature (see, e.g., Lemma 1 of Manski, 1985) that to establish the point identification of the parameter θ in a distribution-free setting, x_t also needs to satisfy certain regularity conditions. Assumption A(c) requires the existence of a relevant, continuous regressor, with large support, which is a standard restriction imposed in MS-type estimators. Assumption A(d) is the familiar full-rank condition. Assumptions A(c) and A(d) are identical to Assumption 2 of Manski (1987).

Assumption A(e) is for scale normalization and parameter space. This is a typical practice for discrete choice models, because the identification of θ is only up to scale. In the semiparametric framework, where no parametric form of $F_{\epsilon|\alpha}$ is specified, identification is often achieved by normalizing the magnitude of the regression coefficients. Assumption A(e) assumes that β is on the unit circle and has a nonzero first element β_1.[9]

HK demonstrate that, if $T \geq 3$ and x_t has time-varying overlap support, θ can be identified under Assumption A.[10] Their proposed approach requires matching all exogenous covariates over time, and results in an estimator with a rate that declines as the number of exogenous covariates increases. The main contribution of this paper is the provision of a set of supplementary conditions, under which the

[9] Our procedure identifies β and γ sequentially, so it is more convenient to normalize the scale of β, rather than that of θ, as in HK.

[10] As is stated in HK, Assumption A is not sufficient for point identifying θ if $T < 3$.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

identification of θ can escape from the necessity of element-by-element matching. Specifically, our approach is based on the following monotonic relationship between a conditional choice probability and an index of the exogenous covariates: For some $s,t \in \mathcal{T}$ such that $t-s \geq 2$,

$$P(y_t = 1|x_s, x_t, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1}, \alpha) \geq P(y_s = 1|x_s, x_t, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1}, \alpha)$$
$$\Leftrightarrow$$
$$x_t'\beta \geq x_s'\beta. \tag{2.2}$$

Note that (2.2) requires that there be at least five ($T \geq 4$) observations per individual observed by the econometrician (i.e., $s = 1$, $t = 3$, and $s+1 = t-1 = 2$). In the simplest case with $T = 4$, (2.2) reduces to $P(y_3 = 1|x_1, x_3, y_0 = y_2 = y_4, \alpha) \geq P(y_1 = 1|x_1, x_3, y_0 = y_2 = y_4, \alpha)$ if and only if $x_3'\beta \geq x_1'\beta$.

Result (2.2) states that the indices $x_s'\beta$ and $x_t'\beta$ rank order the (conditional) probabilities of choosing 1 in periods s and t. To ensure this, conditioning on $y_{s-1} = y_{t-1}$ is obviously necessary. However, as $t-1 > s$, y_s affects y_{t-1} through the dynamics of the model (specifically, through the chain $y_s \to y_{s+1} \to \cdots \to y_{t-1}$). Accordingly, we need to include y_{s+1} in the conditioning set to cut off such state dependence, and further impose the restriction $y_{s+1} = y_{t+1}$ to make the events $\{y_s = 1\}$ and $\{y_t = 1\}$ have symmetric conditioning sets. Only with this symmetry can we invoke time stationarity restrictions on x_t to establish the equivalence in (2.2).

In particular, to reach (2.2), we also need to address the following two concerns. First, x_s (x_t) may affect the value of y_{t+1} (y_{s+1}) via its serial dependence on x_{t+1} (x_{s+1}). Second, the dependence between x_t and y_{t+1} (via x_{t+1}) may change dramatically over time. Both require additional restrictions to be placed on the serial dependence of the stochastic process of x_t. Otherwise, as shown in Appendix A, $x_t'\beta$ is not the unique factor that can rank order the choice probabilities in (2.2).

The following condition, together with Assumption A, is sufficient to establish (2.2), as shown in Appendix A.

Assumption SI. For all $s,t \in \mathcal{T}$ such that $s \neq t$, (a) $x_s \perp x_t|\alpha$, and (b) $x_s \overset{d}{=} x_t|\alpha$.

In Appendix A, we first prove (2.2) under Assumptions A and SI for a special case of model (2.1) with $T = 4$ and $\gamma < 0$. This serves as a roadmap to help readers understand the main ideas. The same arguments can be applied analogously to prove the most general case (see Lemma A.4).

Assumption SI imposes a strong restriction on the dynamic process of the covariate sequence, which requires the process $\{x_t\}$ to be serially independent and strictly stationary, conditional on the individual-specific effects α. In a dynamic fixed effects model, α collects all time-invariant covariates, as well as unobserved individual preferences, abilities, or character traits. In such models, if x_t includes only observed individual characteristics naturally correlated with α, it may be reasonable to further assume that the serial dependence in the process $\{x_t\}$ is also derived from α. Assumption SI implies that we cannot accommodate time

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

trends. We do allow random time effects λ_t that satisfy $(\lambda_s, x_s) \perp (\lambda_t, x_t)|\alpha$, and $(\lambda_s, x_s) \overset{d}{=} (\lambda_t, x_t)|\alpha$.[11]

If x_t contains covariates related to some institutional factors that lead to exogenous variation in, for example, costs of participation, across individuals, Assumption SI may be approximately satisfied by using the differencing, demeaning, or de-trending transformation of these variables. This applies to cases where $\{x_t\}$ exhibits some long-run equilibrium (either deterministic or stochastic trend). The transformed regressor then measures the deviation of x_t from its long-run equilibrium (trend), which, in some cases, might be assumed to be a white noise process affecting the short-run dynamics of the model.[12] Note that such variable transformation may involve model reparameterization. For example, when model (2.1) is the reduced form derived from some structural model, one would need to first reparameterize the structural model accordingly, so that the estimates of the coefficient vector in model (2.1) with transformed covariates can be interpreted in a meaningful way.

HK assume that the support of x_t is overlapping over time, so the differences in the regressors across different time periods have a positive density in a neighborhood of zero. However, evidence presented in Honoré and Tamer (2006) implies that some additional assumption is needed to achieve point identification without performing an element-by-element match, as in HK. Indeed, Assumption SI is the extra condition needed for our approach, compared with the semiparametric estimator in HK.

Under Assumptions A and SI, the identification of θ proceeds in two steps. Proposition 2.1 demonstrates that β can be identified based on moment inequality (2.2), and Proposition 2.2 establishes the identification of γ by matching the value of the index function $x_t'\beta$ in different periods.

PROPOSITION 2.1 (Identification of β). *Assume $T \geq 4$. For all $s, t \in \mathcal{T}$ such that $t \geq s+2$, define*

$$Q_1(b) = \mathbb{E}\{[P(y_t = 1|x_s, x_t, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1}) - P(y_s = 1|x_s, x_t, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1})] \times sgn(x_{ts}'b)\}.$$

Suppose Assumptions A and SI hold. Then, $Q_1(\beta) > Q_1(b)$, for all $b \in \mathcal{B} \setminus \{\beta\}$.

The proof of Proposition 2.1 can be found in Appendix A. Note that our identification strategy for β requires $T = 4$, as a minimum. In this case, $t = s+2$ must hold with $s = 1$ and $t = 3$, and thus $Q_1(b)$ is

$$Q_1(b) = \mathbb{E}\{[P(y_3 = 1|x_1, x_3, y_0 = y_2 = y_4) - P(y_1 = 1|x_1, x_3, y_0 = y_2 = y_4)] \cdot sgn(x_{31}'b)\}. \quad \textbf{(2.3)}$$

[11] We can apply essentially the same arguments used in Appendix A to establish a monotonic relationship analogous to (2.2) for index $\lambda + x_t'\beta$, and identify β and γ similarly.

[12] For example, consider a case where $\{x_t\}$ is a random-walk-plus-drift process (i.e., $x_t = x_0 + a_0 t + \sum_{\tau=1}^{t} e_\tau$). Although $\{x_t\}$ violates Assumption SI, its first differencing $\Delta x_t = x_t - x_{t-1} = a_0 + e_t$ is i.i.d. over time.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

Proposition 2.1 establishes the identification of β, which enables us to identify γ, with β being treated as a known, constant vector. Then, the following proposition shows that γ can be identified by matching the deterministic utility $w_t \equiv x_t'\beta$ in different periods; the proof is presented in Appendix A. Because the key idea for identifying γ uses the insight of Section 4.1 in HK, in what follows, we keep the notation as close to that of HK as possible.

We define the event

$$A = \{y_0 = d_0, \ldots, y_{s-1} = d_{s-1}, y_s = 0, y_{s+1} = d_{s+1}, \ldots, y_{t-1} = d_{t-1}, y_t = 1, \\ y_{t+1} = d_{t+1}, \ldots, y_T = d_T\},$$

and its counterpart

$$B = \{y_0 = d_0, \ldots, y_{s-1} = d_{s-1}, y_s = 1, y_{s+1} = d_{s+1}, \ldots, y_{t-1} = d_{t-1}, y_t = 0, \\ y_{t+1} = d_{t+1}, \ldots, y_T = d_T\},$$

where $d_\tau \in \{0, 1\}$, for $0 \leq \tau \leq T$. Note that y takes the same values other than at time periods (s, t) for A and B: y switches from 0 to 1 at time s and t, respectively, for A, and y switches from 1 to 0 at time s and t, respectively, for B.

We have two cases, based on whether s and t are adjacent. When s and t are adjacent ($t = s + 1$), we define the objective function

$$Q_2(r;\beta) = \mathbb{E}\left\{\left[P(A|x^T, w_t = w_{t+1}) - P(B|x^T, w_t = w_{t+1})\right] \right. \\ \left. \times sgn\left((w_t - w_{t-1}) + r(d_{t+1} - d_{t-2})\right)\right\}.$$

For the case where s and t are not adjacent ($t > s + 1$), we define the objective function

$$\tilde{Q}_2(r;\beta) = \mathbb{E}\{[P(A|x^T, w_{s+1} = w_{t+1}, y_{s+1} = y_{t+1}) - P(B|x^T, w_{s+1} = w_{t+1}, y_{s+1} = y_{t+1})] \\ \times sgn((w_t - w_s) + r(d_{t-1} - d_{s-1}))\}.$$

In the following proposition, we establish the identification of γ by showing that γ uniquely maximizes both $Q_2(r;\beta)$ and $\tilde{Q}_2(r;\beta)$.

From the definitions of A, B, and $Q_2(r;\beta)$, we require $T \geq 3$. In the simplest case when $T = 3$, we have

$$A = \{y_0 = d_0, y_1 = 0, y_2 = 1, y_3 = d_3\} \text{ and } B = \{y_0 = d_0, y_1 = 1, y_2 = 0, y_3 = d_3\},$$

and

$$Q_2(r;\beta) = \mathbb{E}\left\{\left[P(A|x^T, w_2 = w_3) - P(B|x^T, w_2 = w_3)\right] sgn\left((w_2 - w_1) + r(d_3 - d_0)\right)\right\}.$$

$\tilde{Q}_2(r;\beta)$ is not applicable for this case.

Proposition 2.2 (Identification of γ). *Suppose Assumption A holds. We have:*

(i) $Q_2(\gamma;\beta) > Q_2(r;\beta)$ *for all* $r \in \mathcal{R} \setminus \{\gamma\}$.
(ii) $\tilde{Q}_2(\gamma;\beta) > \tilde{Q}_2(r;\beta)$ *for all* $r \in \mathcal{R} \setminus \{\gamma\}$.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

Remark 2.1. When $T \geq 4$, any combination (s,t) of the elements of $\{1,\ldots,T-1\}$ taken two at a time can be used to construct the population objective function to identify γ. For example, in the simplest case $T=4$, feasible choices of (s,t) include $(1,2)$, $(1,3)$, and $(2,3)$. One can use any of these pairs to define the population objective function, either $Q_2(\cdot;\beta)$ or $\tilde{Q}_2(\cdot;\beta)$. Clearly, any one (or combination, e.g., by simply summing them up) of these objective functions can be used to identify γ.

Propositions 2.1 and 2.2 outline a two-step procedure for identifying the preference parameters β and γ, of which Proposition 2.2 uses HK's insight. Note that, as Proposition 2.1 suggests, an additional assumption, SI, enables us to establish the identification of β independently to that of γ in the first step. As a result, it suffices to match the index $x_t'\beta$, rather than each component of x_t over time, as in HK, when identifying γ in the second step. The benefit of doing so is that the two-step procedure avoids the curse of dimensionality caused by matching many explanatory variables (see Theorem 4.1 in Section 4). Our method is particularly competitive when handling high-dimensional models.

The following theorem is an immediate result of Propositions 2.1 and 2.2.

THEOREM 2.1 (Identification of θ). *Suppose $T \geq 4$ and Assumptions A and SI hold. Then, β is identified based on population objective function $Q_1(\cdot)$, and γ is identified based on either population objective function $Q_2(\cdot;\beta)$ or $\tilde{Q}_2(\cdot;\beta)$.*

2.1. Alternative Sufficient Conditions for Identification

Here, we provide an alternative sufficient condition that permits limited dependence of the covariates for the identification. We show in Lemma A.3 that Assumptions A and SD (in below) are sufficient for the inequality in (2.2), which, in turn, shows the identification. In this section, we present and discuss this assumption.

Assumption SD. For all α and $s,t \in \mathcal{T}$,

(a) $f_{\epsilon|\alpha}(\cdot)/F_{\epsilon|\alpha}(\cdot)$ is a nonincreasing function, or equivalently, $f_{\epsilon|\alpha}(\cdot)/[1-F_{\epsilon|\alpha}(\cdot)]$ is a nondecreasing function.
(b) Let $w_t \equiv x_t'\beta$. The joint PDF of w^T conditional on α is exchangeable, that is,

$$f_{w^T|\alpha}(\omega_1,\ldots,\omega_T) = f_{w^T|\alpha}(\omega_{\pi(1)},\ldots,\omega_{\pi(T)})$$

for all permutations $\{\pi(1),\ldots,\pi(T)\}$ defined on the set $\mathcal{T}$.

Assumption SD(a) states that $F_{\epsilon|\alpha}$ has a decreasing inverse Mills ratio, which, together with Assumption SD(b), guarantees the monotonic relation in (2.2), as proved in Appendix A. Assumption SD(a) is satisfied by many common continuous distributions, such as the Gaussian, logistic, Laplace, uniform, gamma,

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

log-normal, Gumbel, and Weibull distributions.[13] However, this property fails if $F_{\epsilon|\alpha}$ has heavy tails (e.g., Student's t-distribution and Cauchy distribution).[14] Note that Assumption SD(a) is a key condition imposed in McFadden (1976) and Silvapulle (1981) for both $-\log F_{\epsilon|\alpha}(\cdot)$ and $-\log(1-F_{\epsilon|\alpha}(\cdot))$ being convex, which guarantees a unique solution for the MLE in cross-sectional models with errors that follow a general distribution.

In model (2.1), the exogenous utility w_t affects the value of y_{t+1} via y_t and its serial dependence with w_{t+1}, conditioning on α. The former is explicitly captured by the coefficient γ. For the latter, Assumption SD(b) restricts the serial dependence of $\{w_t\}$. Assumption SD(b) is weaker than Assumption SI. Under Assumption SI, w_t's are i.i.d. over time, conditional on α. Then, w_t's have an exchangeable joint PDF, as defined in Assumption SD(b). However, the other direction is not always true. As noted in Fox (2007), a common example of an exchangeable PDF for non-independent w_t's is a multivariate normal density with $\mathbb{E}[w_t] = \mu$, $\text{Var}(w_t) = \sigma^2$, and $\text{Corr}(w_s, w_t) = \rho$, for all $s,t \in \{1,\dots,T\}$. More generally, if each w_t can be expressed as $w_t = \varphi(u_t, v)$, where u_t's are i.i.d. random variables, v is a random variable independent of all u_t's, and $\varphi(\cdot,\cdot)$ is some measurable function, then w^T satisfies Assumption SD(b), but not Assumption SI. Similar exchangeability assumptions are imposed in Altonji and Matzkin (2005) and Chen, Khan, and Tang (2018).

A few remarks are in order about how our identification conditions are related to the existing literature.

Remark 2.2. Compared with HK, our approach relies on additional assumptions restricting the serial dependence of strictly exogenous regressors x_t and requires $T \geq 4$. These conditions make identification without element-by-element matching of x_t possible. HK construct identifying inequalities similar to our (2.2). To obtain point identification, HK use probabilities of specific sequences of y^T conditional on event $\{x_s = x_t\}$, for some $s,t \in \{1,\dots,T\}$. Instead, our approach matches y_t in different time periods to construct identifying inequalities, allowing for point identification without element-by-element matching of x_t. However, as discussed after the introduction of our key identifying condition (2.2), x_t can affect the choice probabilities in (2.2) through the utility index $x_t'\beta$ and its serial dependence with x in other time periods. Assumptions SI and SD restrict this dependence, ensuring that choice probabilities are solely rank-ordered by $x_t'\beta$. Additionally, our approach requires comparing $x_t'\beta$ in non-adjacent time periods

[13] In a mixture model, for example,

$$f_{\epsilon|\alpha}(e) = \sum_{m=1}^{M} \pi_m f_{\epsilon|\alpha}(e;\vartheta_m)$$

with mixing proportions π_m, $\sum_{m=1}^{M}\pi_m = 1$, where each component density has a different parameter vector ϑ_m, Assumption SD(a) holds for $F_{\epsilon_t|\alpha}(\cdot)$ if it is satisfied by all component distributions $F_{\epsilon|\alpha}(\cdot;\vartheta_m)$.

[14] More precisely, Assumption SD(a) does not hold globally for these distributions. For example, it is not difficult to find that this assumption holds for Student's t and Cauchy on $[-L,\infty)$ for some positive L.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

(i.e., $t > s+1$ in (2.2)). As a result, we need observation of one more period for each individual, compared with HK's method, which can achieve point identification using $x_t'\beta$ in adjacent time periods ($t = s+1$).

Remark 2.3. Second, our identification conditions are non-nested with those in the literature, assuming exclusion restrictions, such as Honoré and Lewbel (2002), Chen et al. (2018, 2019), and Williams (2019). Chen et al. (2019) show that Honoré and Lewbel (2002) essentially require the serial independence of the excluded regressor. Williams (2019) requires that the other strictly exogenous regressors are conditionally independent of the past values of the excluded regressor. In addition to specific restrictions on the dynamic process for the covariates, the identification results of these studies rely on the existence of at least one "excluded regressor" conditionally independent of the individual fixed effects α. Conversely, our approach allows for arbitrary correlation between x_t and α.

3. ESTIMATION

Applying the analogy principle, the identification results presented in Section 2 can be translated into a two-step estimation procedure. In the first step, we obtain an MS estimator (with binary weights) $\hat{\beta}$ of β. In the second step, γ is estimated by a localized MS procedure matching the estimated index $x_t'\hat{\beta}$ over time. Each of the two steps is described, in turn, below.

In Sections 3.1 and 3.2, we restrict our discussion to the model with $T = 4$ to streamline exposition in subsequent sections. The same method can be applied, with straightforward modification, to models with longer panels. We provide objective functions for general cases with $T \geq 4$ in Section 3.3.

3.1. Estimation of β with $T = 4$

Assuming a random sample of n individuals, we propose the following weighted MS estimator $\hat{\beta}$ of β, defined as the maximizer over the parameter space $\mathcal{B}$:

$$\hat{\beta} = \arg\max_{b \in \mathcal{B}} Q_{1n}(b), \tag{3.1}$$

where

$$Q_{1n}(b) = \frac{1}{n}\sum_{i=1}^{n} 1[y_{i0} = y_{i2} = y_{i4}](y_{i3} - y_{i1}) \cdot sgn(x_{i31}'b). \tag{3.2}$$

Because we restrict the search within a compact set $\mathcal{B}$ and the objective function (3.2) is bounded and continuous, the maximizer $\hat{\beta}$ of the maximization problem (3.1) does exist. However, $\hat{\beta}$ may not be unique because the objective function (3.2) is essentially a step function for any finite samples. Nonetheless, as guaranteed by Theorem 4.1 in Section 4, $\hat{\beta}$ is in a small neighborhood of the true parameter β with a high probability for a sufficiently large sample size.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

It is clear from expression (3.2) that only observations that satisfy $y_{i1} \neq y_{i3}$, $y_{i0} = y_{i2}$, and $y_{i2} = y_{i4}$ are used in the estimation. That is, the objective function uses only "switchers" with choice changes in periods 1 and 3, with the same choices in their previous and subsequent periods, respectively. This feature reduces the "effective" sample size for the estimator of β. HK's estimator has a similar problem, because it also uses only switchers, and needs to match x_t over time. Our estimator (3.1) is more applicable when the model has many regressors, especially discrete regressors that must be matched exactly over time when applying HK's procedure.

3.2. Estimation of γ with $T = 4$

Proposition 2.2 motivates a localized MS estimator $\hat{\gamma}$ of γ, defined here as the maximizer over the parameter space $\mathcal{R}$ of the objective function[15]

$$Q_{2n}(r;\beta) = \frac{1}{n}\sum_{i=1}^{n}\{1[x_{i2}'\beta = x_{i3}'\beta](y_{i2} - y_{i1}) \cdot sgn(x_{i21}'\beta + r(y_{i3} - y_{i0})) + 1[x_{i3}'\beta = x_{i4}'\beta](y_{i3} - y_{i2}) \cdot sgn(x_{i32}'\beta + r(y_{i4} - y_{i1}))\}. \quad (3.3)$$

Expression (3.3) is the sample analog of $Q_2(r;\beta)$ in Proposition 2.2 after taking the union of events A and B for all possible values of $d_0, d_1, \ldots, d_4$. As with objective function (3.2), (3.3) also uses only data on switchers (i.e., satisfying $A \cup B$) who make different choices in the two periods compared. In addition, (3.3) also requires a match in $x_t'\beta$.

Note that this estimator is not feasible, because β is unknown, and it is of probability zero to have exactly matched indices ($x_{is}'\beta = x_{it}'\beta$) in the presence of continuous regressors. To resolve the first concern, we propose replacing the unknown parameter β in expression (3.3) with the $\hat{\beta}$ obtained from (3.1), which is shown to be (cube-root n) consistent in Section 4.

For the second concern, we use kernel weights

$$\mathcal{K}_{h_n}((x_{it} - x_{is})'b), \text{ for all } s,t \in \mathcal{T} \text{ and } b \in \mathcal{B},$$

instead of $1[x_{is}'b = x_{it}'b]$. $\mathcal{K}_{h_n}(\cdot)$ is defined as $h_n^{-1}\mathcal{K}(\cdot/h_n)$, where $\mathcal{K}(\cdot)$ is a kernel density function and h_n is a bandwidth sequence that converges to zero as $n \to \infty$. The idea is to replace the binary weights for $x_{is}'\hat{\beta} = x_{it}'\hat{\beta}$ with weights that depend inversely on the magnitude of $(x_{it} - x_{is})'\hat{\beta}$, giving more weight to observations with $(x_{it} - x_{is})'\hat{\beta}$ closer to zero. We discuss the choice of the tuning parameter h_n with illustrating examples in Section 6.2.

Then, we propose the following kernel-weighted MS estimator $\hat{\gamma}$ of γ:

$$\hat{\gamma} = \arg\max_{r \in \mathcal{R}} Q_{2n}^K(r;\hat{\beta}), \quad (3.4)$$

[15] If one knew β, the estimation of γ requires only $T = 3$, that is, using the first line of (3.3) as the objective function.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

where

$$Q_{2n}^{K}(r;\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{\mathcal{K}_{h_n}(x_{i32}'\hat{\beta})(y_{i2}-y_{i1})\cdot sgn(x_{i21}'\hat{\beta}+r(y_{i3}-y_{i0})) + \mathcal{K}_{h_n}(x_{i43}'\hat{\beta})(y_{i3}-y_{i2})\cdot sgn(x_{i32}'\hat{\beta}+r(y_{i4}-y_{i1}))\}. \quad (3.5)$$

Remark 3.1. Note that objective function (3.5) is associated with the population objective function $Q_2(r;\beta)$ in Proposition 2.2, which uses only observations of adjacent time periods. Applying the same idea to the population objective function $\tilde{Q}_2(r;\beta)$ yields the following objective function, using observations that are not adjacent:

$$\tilde{Q}_{2n}^{K}(r;\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}1[y_{i2}=y_{i4}]\mathcal{K}_{h_n}(x_{i42}'\hat{\beta})(y_{i3}-y_{i1})\cdot sgn(x_{i31}'\hat{\beta}+r(y_{i2}-y_{i0})).$$

In practice, to make full use of all observations, one can consider using $Q_{2n}^{K}(r;\hat{\beta})+\tilde{Q}_{2n}^{K}(r;\hat{\beta})$ as the objective function for the estimation of γ.

Remark 3.2. Calculating the MS-type of estimators (3.1) and (3.4) is challenging, as it is for the semiparametric estimator of HK. Following Fox (2007) and Yan and Yoo (2019), we suggest using a global optimization method called the *differential evolution* (DE) algorithm. The DE algorithm, introduced by Storn and Price (1997), is specifically designed to search the global optimum of a real-valued function with real-valued parameters. Notably, it does not require the objective function to be continuous or differentiable. The DE algorithm has been widely used in engineering applications, and its performance as a global optimization algorithm has been studied extensively (see, e.g., Price, Storn, and Lampinen, 2006). To implement the DE algorithm, one can use the "`DEoptim`" package in R.[16] The following statement is quoted from the R documentation for the "`DEoptim`" package, which provides a brief introduction. Interested readers are advised to consult this documentation for more information on the implementation and usage of this algorithm.

"*Differential Evolution (DE) is a search heuristic introduced by Storn and Price (1997). Its remarkable performance as a global optimization algorithm on continuous numerical minimization problems has been extensively explored; see Price et al. (2006). DE belongs to the class of genetic algorithms which use biology-inspired operations of crossover, mutation, and selection on a population in order to minimize an objective function over the course of successive generations. As with other evolutionary algorithms, DE solves optimization problems by evolving a population of candidate solutions using alteration and selection operators. DE uses floating-point instead of bit-string encoding of population members, and arithmetic operations instead of logical operations in mutation. DE is particularly well-suited to find the global optimum of a real-valued function of*

[16] https://cran.r-project.org/web/packages/DEoptim/index.html.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

real-valued parameters, and does not require that the function be either continuous or differentiable."

Note that the 2SMS procedure described in (3.1), (3.2), (3.4), and (3.5) does not require matching each covariate in x_{it} over time, as it does in HK. As a result, the rates of convergence of $\hat{\beta}$ and $\hat{\gamma}$ are independent of the number of continuous covariates in x_{it}, in contrast to the procedure of HK. In view of existing results on the MS estimators (e.g., Manski, 1985, 1987; Kim and Pollard, 1990; Seo and Otsu, 2018), we expect the limiting distributions of $\hat{\beta}$ and $\hat{\gamma}$ to be non-Gaussian and their rates of convergence to be $O_P(n^{-1/3})$ and $O_P((nh_n)^{-1/3})$, respectively. Section 4 states sufficient conditions under which these asymptotic properties can be derived.

3.3. Estimation with $T \geq 4$

A longer panel allows for more objective functions of similar form. Collectively, these objective functions (by, e.g., summing them) can be used to obtain more accurate estimates of θ for finite samples. For the case with $T \geq 4$, estimators for β and γ that best use the data can be obtained as follows. For β, we find $\hat{\beta}$ by maximizing

$$Q_{1n}(b) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t>s+1} 1[y_{is-1} = y_{it-1}]1[y_{is+1} = y_{it+1}](y_{it} - y_{is})sgn\left((x_{it} - x_{is})'b\right).$$

Once $\hat{\beta}$ is obtained, we estimate γ by maximizing

$$Q_{2n}^{K}(r;\hat{\beta}) + \tilde{Q}_{2n}^{K}(r;\hat{\beta})$$

with respect to r, where

$$Q_{2n}^{K}(r;\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=2}^{T-1}\mathcal{K}_{h_n}((x_{it+1} - x_{it})'\hat{\beta})(y_{it} - y_{it-1})sgn((x_{it} - x_{it-1})'\hat{\beta} + r(y_{it+1} - y_{it-2}))$$

is for the case with $t = s+1$, and

$$\tilde{Q}_{2n}^{K}(r;\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{T-3}\sum_{t=s+2}^{T-1}\{1[y_{is+1} = y_{it+1}]\mathcal{K}_{h_n}((x_{it+1} - x_{is+1})'\hat{\beta}) \times (y_{it} - y_{is})sgn((x_{it} - x_{is})'\hat{\beta} + r(y_{it-1} - y_{is-1}))\}$$

is for the case with $t > s+1$.

4. ASYMPTOTIC PROPERTIES

The estimators proposed in Section 3 are of the same structure and differ only in that they each use a different fraction of observations in the sample. We expect that they have similar asymptotic properties. Therefore, it suffices to show the asymptotics for the estimators in Sections 3.1 and 3.2, for the case $T = 4$.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

The asymptotic properties of the estimators in Section 3.3 can be derived in a similar way.

As is standard in the literature, such as Kim and Pollard (1990), we start the analysis by introducing modified objective functions for $\hat{\beta}$ and $\hat{\gamma}$. The new objective functions are monotone (linear) transformations of (3.2) and (3.5), respectively. As a result, working with them does not change the values of $\hat{\beta}$ and $\hat{\gamma}$, but can facilitate the derivation process.

Because adding terms not related to b will not affect the optimization over b, and $1[a > 0] = (sgn(a) + 1)/2$ for all $a \in \mathbb{R}$, $\hat{\beta}$ obtained from the following objective function is identical to that from (3.2),

$$\hat{\beta} = \arg\max_{b \in \mathcal{B}} n^{-1} \sum_{i=1}^{n} \xi_i(b),$$

where

$$\xi_i(b) \equiv 1[y_{i0} = y_{i2} = y_{i4}](y_{i3} - y_{i1}) \left(1\left[x'_{i31} b > 0\right] - 1\left[x'_{i31}\beta > 0\right]\right). \tag{4.1}$$

For the same reason, $\hat{\gamma}$ can be obtained equivalently from

$$\hat{\gamma} = \arg\max_{r \in \mathcal{R}} n^{-1} \sum_{i=1}^{n} \varsigma_{ni}(r, \hat{\beta}),$$

where

$$\begin{aligned} \varsigma_{ni}(r, \hat{\beta}) \equiv\ & \mathcal{K}_{h_n}\left(x'_{i32}\hat{\beta}\right)(y_{i2} - y_{i1})\, 1\left[x'_{i21}\hat{\beta} + r(y_{i3} - y_{i0}) > 0\right] \\ & + \mathcal{K}_{h_n}\left(x'_{i43}\hat{\beta}\right)(y_{i3} - y_{i2})\, 1\left[x'_{i32}\hat{\beta} + r(y_{i4} - y_{i1}) > 0\right]. \end{aligned} \tag{4.2}$$

The following technical assumptions are needed for the asymptotics of $\hat{\beta}$ and $\hat{\gamma}$.

Assumption 1. The vectors $\left(x_i^T, y_i^T, y_{i0}\right)'$ with $T \geq 4$ are i.i.d. across individuals.

Assumption 2. $n^{-1}\sum_{i=1}^{n} \xi_i(\hat{\beta}) \geq \max_{b \in \mathcal{B}} n^{-1}\sum_{i=1}^{n} \xi_i(b) - o_P\left(n^{-2/3}\right)$ and $n^{-1}\sum_{i=1}^{n} \varsigma_{ni}(\hat{\gamma}, \hat{\beta}) \geq \max_{r \in \mathcal{R}} n^{-1}\sum_{i=1}^{n} \varsigma_{ni}(r, \hat{\beta}) - o_P\left((nh_n)^{-2/3}\right)$.

Assumption 3. The joint density function for α, covariates x^T, and ϵ^T are continuously differentiable. The density function and its first-order derivatives are uniformly bounded. Further,

$$f\left(\epsilon_t | \alpha, x^T, \epsilon_1, \ldots, \epsilon_{t-1}, \epsilon_{t+1}, \ldots, \epsilon_T\right) \text{and } f\left(x_t | \alpha, x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T, \epsilon^T\right)$$

are continuous differentiable with respect to all arguments. The conditional densities and their derivatives are uniformly bounded.

Assumption 4. The kernel function $\mathcal{K}(u)$ is nonnegative, symmetric about zero, continuous differentiable, has compact support, and satisfies $\int_{\mathbb{R}} \mathcal{K}(u)\, du = 1$.

Assumption 5. $h_n \to 0, nh_n \to \infty,$ and $nh_n^4 \to 0$ as $n \to \infty$.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

Assumption 2 is standard in the literature and precisely defines our estimator. Assumption 3 is for technical convenience; it ensures certain functions defined in the proof of Theorem 4.1 are differentiable, so that V_1 and V_2 (defined in Theorem 4.1) have simple representations. In the case of discrete explanatory variables which violates Assumption 3, V_1 and V_2 can be shown to be well defined, but with more tedious calculations and notation. Assumption 4 collects some standard restrictions on kernel functions. The symmetry of $\mathcal{K}(u)$ ensures that the bias term is of the order h_n^2. In Assumption 5, $nh_n \to \infty$ is standard, and $nh_n^4 \to 0$ ensures the bias term from the kernel estimation is asymptotically negligible.

THEOREM 4.1. *Suppose $T \geq 4$ and Assumptions A, SI (or SD), and 1–5 hold. Then,*

1. $\hat{\beta} - \beta = O_P\left(n^{-1/3}\right)$, *and*

$$n^{1/3}(\hat{\beta} - \beta) \xrightarrow{d} \arg\max_{s \in \mathbb{R}^K} Z_1(\boldsymbol{s}),$$

where $Z_1(\boldsymbol{s})$ is a Gaussian process with continuous sample paths, expected value $\frac{1}{2}\boldsymbol{s}'V_1\boldsymbol{s}$, and covariance kernel $H_1(\boldsymbol{s},\boldsymbol{t})$. V_1 and H_1 are defined in expressions (B.1) and (B.2), respectively.

2. $\hat{\gamma} - \gamma = O_P((nh_n)^{-1/3})$, *and*

$$(nh_n)^{1/3}\left(\hat{\gamma} - \gamma\right) \xrightarrow{d} \arg\max_{s \in \mathbb{R}} Z_2(s),$$

where $Z_2(s)$ is a Gaussian process with continuous path, expected value $\frac{1}{2}V_2 s^2$, and covariance kernel $H_2(s,t)$. V_2 and H_2 are defined in expressions (B.3) and (B.4), respectively.

Kim and Pollard (1990) and Seo and Otsu (2018) derive the cube-root asymptotics for a class of estimators by means of empirical processes. For a comprehensive treatment of this technique, see van der Vaart and Wellner (1996). Our estimators fall into this category. In particular, they are more closely related to Seo and Otsu (2018). The main body of the proof for Theorem 4.1 verifies the technical conditions in Seo and Otsu (2018), applies their asymptotics results to our estimators, and calculates the technical terms needed for the asymptotics such as V_1, H_1, V_2, and H_2.

Note that the asymptotics of $\hat{\gamma}$ are the same as in the case where the true value of β is used. Intuitively, $\hat{\beta}$ converges to β faster than $\hat{\gamma}$ does to γ, and the objective function (4.2), after proper normalization, uniformly converges to the limit over a compact set of $\left(b', r\right)'$ around $\left(\beta', \gamma\right)'$. The details can be found in the proof of Theorem 4.1, which is presented in Appendix B.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

5. INFERENCE

The asymptotic distributions of $\hat{\beta}$ and $\hat{\gamma}$ are complicated and do not have an analytical form. As a result, inference using the asymptotic distribution directly is difficult to implement. One may consider smoothing the objective functions, in the spirit of Horowitz (1992), to attain faster rates of convergence and asymptotic normality.[17] However, this requires selecting additional kernel functions and tuning parameters, and then computing consistent estimates for asymptotic variances. As an alternative, we seek to use more direct sampling methods (e.g., bootstrap). Unfortunately, Abrevaya and Huang (2005) have proved the inconsistency of the classic bootstrap for the MS estimators. We expect that the classic bootstrap does not work for our estimators either.

For the ordinary MS estimator, valid inference can be conducted using subsampling (Delgado, Rodríguez-Poo, and Wolf, 2001), m-out-of-n bootstrap (Lee and Pun, 2006), the numerical bootstrap (Hong and Li, 2020), and a model-based bootstrap procedure that analytically modifies the criterion function (Cattaneo, Jansson, and Nagasawa, 2020), among other procedures.[18] These methods, with certain modifications, can be justified to be valid for our estimators.

Monte Carlo evidence demonstrated in Hong and Li (2020) suggests that their proposed approach outperforms the subsampling and the m-out-of-n bootstrap in finite samples. Based on these results, we focus on the numerical bootstrap. We provide a brief discussion on the classic bootstrap and the m-out-of-n bootstrap in Appendix E of the Supplementary Material.[19]

We next introduce some additional notation. Let $(y_j^{T*\prime}, x_j^{T*\prime})', j = 1, \ldots, n$, be a random sample drawn with replacement from the collection of the sample values $\left(y_1^{T\prime}, x_1^{T\prime}\right)', \left(y_2^{T\prime}, x_2^{T\prime}\right)', \ldots, \left(y_n^{T\prime}, x_n^{T\prime}\right)'$. Let $\xi_j^*(b)$ denote $\xi(b)$ evaluated at $(y_j^{T*\prime}, x_j^{T*\prime})'$, specifically,

$$\xi_j^*(b) \equiv 1\left[y_{j0}^* = y_{j2}^* = y_{j4}^*\right]\left(y_{j3}^* - y_{j1}^*\right)\left(1\left[x_{j31}^{*\prime} b > 0\right] - 1\left[x_{j31}^{*\prime}\beta > 0\right]\right).$$

Similarly, we define $\varsigma_{nj}^*(r, b)$ as

$$\varsigma_{nj}^*(r,b) \equiv \mathcal{K}_{h_n}\left(x_{j32}^{*\prime} b\right)\left(y_{j2}^* - y_{j1}^*\right)\left(1\left[x_{j21}^{*\prime} b + r\left(y_{j3}^* - y_{j0}^*\right) > 0\right] - 1\left[x_{j21}^{*\prime}\beta + \gamma\left(y_{j3}^* - y_{j0}^*\right) > 0\right]\right)$$

$$+ \mathcal{K}_{h_n}\left(x_{j43}^{*\prime} b\right)\left(y_{j3}^* - y_{j2}^*\right)\left(1\left[x_{j32}^{*\prime} b + r\left(y_{j4}^* - y_{j1}^*\right) > 0\right] - 1\left[x_{j32}^{*\prime}\beta + \gamma\left(y_{j4}^* - y_{j1}^*\right) > 0\right]\right).$$

[17] See also Kyriazidou (1997) and Charlier (1997).

[18] The case-specific, smooth bootstrap method proposed by Patra, Seijo, and Sen (2018) is also valid for the MS estimator of Manski (1975, 1985). However, this method is difficult to generalize to our case.

[19] We show in Appendix E of the Supplementary Material that the classic bootstrap is not consistent for our estimators (Appendix E.2 of the Supplementary Material), while the m-out-of-n bootstrap is still valid (Appendix E.3 of the Supplementary Material). Note that we re-use some notation in this appendix for notational convenience. To avoid confusion, all notation in each subsection is specific to the procedure discussed in that subsection.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

5.1. Numerical Bootstrap

Hong and Li (2020) develop a numerical bootstrap procedure for cases in which the classic bootstrap does not work. They demonstrate that their method works for a class of M-estimators that converge at rate n^a for some $a \in (1/4, 1]$. The estimator $\hat{\beta}$ proposed in Section 3 fits their framework directly, but $\hat{\gamma}$ does not. With a slight modification of their proof, we show that the numerical bootstrap also works for $\hat{\gamma}$.

The numerically bootstrapped $\hat{\beta}^*$ and $\hat{\gamma}^*$ are constructed from

$$\hat{\beta}^* = \arg\max_{b \in \mathcal{B}} \left\{ n^{-1} \sum_{i=1}^{n} \xi_i(b) + (n\varepsilon_n)^{1/2} \cdot n^{-1} \sum_{j=1}^{n} \left(\xi_j^*(b) - n^{-1} \sum_{i=1}^{n} \xi_i(b) \right) \right\} \tag{5.1}$$

and

$$\hat{\gamma}^* = \arg\max_{r \in \mathcal{R}} \left\{ n^{-1} \sum_{i=1}^{n} \varsigma_{ni}\left(r, \hat{\beta}\right) + (n\varepsilon_n)^{1/2} \cdot n^{-1} \sum_{j=1}^{n} \left(\varsigma_{nj}^*\left(r, \hat{\beta}\right) - n^{-1} \sum_{i=1}^{n} \varsigma_{ni}\left(r, \hat{\beta}\right) \right) \right\}, \tag{5.2}$$

where $\varepsilon_n \to 0, n\varepsilon_n \to \infty$, and $(y_j^{T*\prime}, x_j^{T*\prime})', j = 1, \ldots, n$, are drawn independently from the collection of the sample values $\left(y_1^{T\prime}, x_1^{T\prime}\right)', \left(y_2^{T\prime}, x_2^{T\prime}\right)', \ldots, \left(y_n^{T\prime}, x_n^{T\prime}\right)'$, with replacement. ε_n^{-1} plays a similar role as the m in the m-out-of-n bootstrap procedure. For $\hat{\gamma}^*$, we additionally require $\varepsilon_n^{-1} h_n \to \infty$ and $\varepsilon_n^{-1} h_n^4 \to 0$. Following the same arguments as in the discussion below (3.2) for $\hat{\beta}$, the maximizer $\hat{\beta}^*$ exists, but its uniqueness is not guaranteed due to the non-smoothness of its step objective function. The second term in (5.1) can be shown to be $o_P(1)$. The first term, sharing a similar structure to the objective function (3.2), dominates in (5.1). As a result, we anticipate $\hat{\beta}^*$ to be close to β as $n \to \infty$. Similar arguments apply to $\hat{\gamma}^*$.

We claim that

$$\varepsilon_n^{-1/3} \left(\hat{\beta}^* - \hat{\beta} \right) \xrightarrow{d} \arg\max_{s \in \mathbb{R}^K} \left(\frac{1}{2} \boldsymbol{s}' V_1 \boldsymbol{s} + W_1(\boldsymbol{s}) \right)$$

and

$$\left(\varepsilon_n^{-1} h_n\right)^{1/3} \left(\hat{\gamma}^* - \hat{\gamma} \right) \xrightarrow{d} \arg\max_{s \in \mathbb{R}} \left(\frac{1}{2} V_2 s^2 + W_2(s) \right),$$

where W_1 and W_2 are mean zero Gaussian processes with covariance kernels H_1 and H_2, respectively. An outline of the proof of why the numerical bootstrap works and the way to modify the proof in Hong and Li (2020) to accommodate $\hat{\gamma}$ is provided in Appendix E.1 of the Supplementary Material.

5.2. Procedures in Details

We investigate the finite-sample properties of the bootstrap method discussed in Section 5.1 using Monte Carlo experiments in Section 6, and defer the discussion

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

on the choices of their tuning parameters to Section 6.2. Here, we provide the algorithm for constructing the 95% confidence intervals (CIs) for β and γ.

The numerical bootstrap proceeds as follows:

1. Draw $(y_j^{T*\prime}, x_j^{T*\prime})', j = 1, \ldots, n,$ independently, with replacement, from the original sample.
2. Obtain $\hat{\beta}^*$ and $\hat{\gamma}^*$ from equations (5.1) and (5.2).
3. Repeat Steps 1 and 2 for B times independently and obtain a sequence of $(\hat{\beta}^*, \hat{\gamma}^*)$, say, $\{(\hat{\beta}^{*(b)}, \hat{\gamma}^{*(b)})\}_{b=1}^B$.
4. Let $Q_{\hat{\beta}^*}(\tau)$ denote the τth quantile of $\{\hat{\beta}^{*(b)}\}_{b=1}^B, 0 \leq \tau \leq 1$. Define $Q_{\hat{\gamma}^*}(\tau)$ analogously. The 95% CIs for β and γ are constructed, respectively, as

$$\left[\hat{\beta} - n^{-1/3} \cdot \varepsilon_n^{-1/3}(Q_{\hat{\beta}^*}(0.975) - \hat{\beta}),\ \hat{\beta} - n^{-1/3} \cdot \varepsilon_n^{-1/3}(Q_{\hat{\beta}^*}(0.025) - \hat{\beta})\right]$$

and

$$\left[\hat{\gamma} - n^{-1/3} \cdot \varepsilon_n^{-1/3}\left(Q_{\hat{\gamma}^*}(0.975) - \hat{\gamma}\right),\ \hat{\gamma} - n^{-1/3} \cdot \varepsilon_n^{-1/3}\left(Q_{\hat{\gamma}^*}(0.025) - \hat{\gamma}\right)\right].$$

6. MONTE CARLO EXPERIMENTS

6.1. Simulation Setup

In this section, we investigate the finite-sample performance of the proposed estimators by means of Monte Carlo experiments. We start by considering a benchmark design similar to that used in HK. Specifically, this design (referred to as Design 1 hereafter) is specified as follows:

$$y_{i0} = 1\left[\beta_1 x_{i0,1} + \beta_2 x_{i0,2} + \alpha_i - \epsilon_{i0} > 0\right],$$
$$y_{it} = 1\left[\beta_1 x_{it,1} + \beta_2 x_{it,2} + \gamma y_{it-1} + \alpha_i - \epsilon_{it} > 0\right],\ \ t \in \{1,2,3,4\},$$

where:

- $\beta \equiv (\beta_1, \beta_2)' = (1,1)'$ and $\gamma = -1$,
- $x_{it,j} = \frac{\sqrt{15}}{4} u_{it,j} + \frac{1}{4} u_{it,3}, j = 1,2,\ \left(u_{it,1}, u_{it,2}, u_{it,3}\right) \overset{d}{\sim} N(0_{3\times 1}, I_{3\times 3})$, and $\left(u_{it,1}, u_{it,2}, u_{it,3}\right)$ are i.i.d. across i and t,[20]
- $\alpha_i = \left(x_{i0,2} + x_{i1,2} + x_{i2,2} + x_{i3,2} + x_{i4,2}\right)/5$,
- $\epsilon_{it} \overset{d}{\sim} \left(\pi^2/3\right)^{-1/2} \cdot \text{Logistic}(0,1)$ and are i.i.d. across i and t, and
- $\left(u_{\cdot,1}, u_{\cdot,2}, u_{\cdot,3}\right)$ and $\epsilon_{\cdot}$ are independent of each other.

In the second design (hereafter, Design 2), the model and the coefficients are the same as in Design 1, but $x_{\cdot,1}$ and $x_{\cdot,2}$ are autocorrelated over time. Specifically, we have:

[20] Note that all covariates are correlated with each other and have a standard deviation of one.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

– $x_{i0,j} = \frac{\sqrt{15}}{4} u_{i0,j} + \frac{1}{4} u_{i0,3}, j = 1,2,$ and $x_{it,j} = \frac{1}{2} x_{it-1,j} + \frac{\sqrt{3}}{2} \left(\frac{\sqrt{15}}{4} u_{it,j} + \frac{1}{4} u_{it,3} \right)$, $j = 1,2$ for all $t \geq 1$, where $\left(u_{it,1}, u_{it,2}, u_{it,3}\right) \overset{d}{\sim} N\left(0_{3\times 1}, I_{3\times 3}\right)$ and $\left(u_{it,1}, u_{it,2}, u_{it,3}\right)$ are i.i.d. across i and t,
– $\left(u_{\cdot,1}, u_{\cdot,2}, u_{\cdot,3}\right)$ and $\epsilon_{\cdot}$ are independent of each other.

Note that the setup of Design 2 violates both Assumption SI and the exchangeability condition stated in Assumption SD. We conduct this Monte Carlo study to develop insight into the practical consequences of the failure of these sufficient (but not necessary) conditions. That is, we examine the extent to which serial dependence in exogenous covariates may affect the identification.

In the third to fifth designs (Designs 3–5, respectively), the setup is the same as that in Design 1, except that we add one, two, and three more covariates (in Designs 3–5, respectively) to examine how our estimators perform in higher-dimensional designs. Specifically, in Design $k, k = 3,4,$ and 5,

$$y_{i0} = 1\left[\beta_1 x_{i0,1} + \beta_2 x_{i0,2} + \cdots + \beta_k x_{i0,k} + \alpha_i - \epsilon_{i0} > 0\right],$$
$$y_{it} = 1\left[\beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_k x_{it,k} + \gamma y_{it-1} + \alpha_i - \epsilon_{it} > 0\right], \quad t \in \{1,2,3,4\},$$

where:

– $\beta \equiv (\beta_1, \beta_2, \ldots, \beta_k)' = (1,1,\ldots,1)'$ and $\gamma = -1$,
– $x_{it,j} = \frac{\sqrt{15}}{4} u_{it,j} + \frac{1}{4} u_{it,k+1}, j = 1,2,\ldots,k,$ $\left(u_{it,1}, u_{it,2}, \ldots, u_{it,k+1}\right) \overset{d}{\sim} N\left(0_{(k+1)\times 1}, I_{(k+1)\times(k+1)}\right)$, and $\left(u_{it,1}, u_{it,2}, \ldots, u_{it,k+1}\right)$ are i.i.d. across i and t,
– $\alpha_i = \left(x_{i0,2} + x_{i1,2} + x_{i2,2} + x_{i3,2} + x_{i4,2}\right)/5$,
– $\epsilon_{it} \overset{d}{\sim} \left(\pi^2/3\right)^{-1/2} \cdot \text{Logistic}\,(0,1)$ and are i.i.d. across i and t, and
– $\left(u_{\cdot,1}, u_{\cdot,2}, \ldots, u_{\cdot,k+1}\right)$ and $\epsilon_{\cdot}$ are independent of each other.

We also explore the impact of the serial dependence of x_{it} on the estimation and inference for the models in Designs 3–5 in Appendix F of the Supplementary Material. We adopt a similar method to Design 2 for this analysis, which offers additional insights into the robustness and performance of our proposed method.

For the estimation of β, we adopt the objective function (3.2). To estimate γ, we use the objective function (3.5) with the Epanechnikov kernel function. That is,

$$\mathcal{K}(u) = \frac{3}{4}\left(1 - u^2\right) 1\left[|u| \leq 1\right],$$

which satisfies Assumption 4 with a compact support. We discuss the choice of bandwidth sequence h_n in Section 6.2.

For inference, we investigate the finite-sample performance of the numerical bootstrap (Section 5.1). The 95% CIs are obtained from $B = 199$ independent draws and estimations (see Section 5.2 for the details of the implementation).

Recall that only the observations with $\{y_{i0} = y_{i2} = y_{i4}$ and $y_{i1} \neq y_{i3}\}$ are used to estimate β. In all designs, the effective observations, which are useful for

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

estimating β, comprise about 14% of the whole sample. Similarly, for γ, only observations with either $\{y_{i1} \neq y_{i2}$ and $y_{i0} \neq y_{i3}\}$ or $\{y_{i2} \neq y_{i3}$ and $y_{i1} \neq y_{i4}\}$ are useful. In all designs, about 31% to 39% of the observations are effective for γ. For each design, we consider sample sizes of 2,500, 5,000, 10,000, and 20,000. All the estimation and inference results (based on 199 draws and estimation) presented in this section are based on 1,000 replications of each design and each sample size.

Furthermore, we compare our method with the parametric (Logit) and semi-parametric (distribution-free) estimators of HK. We use the objective functions linked to these two estimators, as defined in Section 4.1 of HK (for panel data where $T > 3$), to implement their methods. To facilitate the comparison, we apply the same scale normalization, specified in Assumption A(e), to both HK's two estimators and our own. Specifically, we normalize the vector of β's to have a Euclidean norm of one. Note that HK's Logit estimator does not require scale normalization for the preference coefficients, because it assumes that the error terms follow a standard logistic distribution. Therefore, if we choose to apply scale normalization to β, we should also estimate a scale parameter in the Logit model to regain one degree of freedom in the parameter space. For example, for two adjacent time periods t and $t+1$ in Designs 1 and 2, the log-likelihood function is written as

$$\sum_{i=1}^{n} 1\left[y_{it} + y_{it+1} = 1\right] \sigma_n^{-2} \mathcal{K}\left(\frac{x_{it+1,1} - x_{it+2,1}}{\sigma_n}\right) \mathcal{K}\left(\frac{x_{it+1,2} - x_{it+2,2}}{\sigma_n}\right)$$
$$\times \log\left(\frac{\exp\left(\left[\left(x_{it,1} - x_{it+1,1}\right) b_1 + \left(x_{it,2} - x_{it+1,2}\right) b_2 + r\left(y_{it-1} - y_{it+2}\right)\right]/s\right)^{y_{it}}}{1 + \exp\left(\left[\left(x_{it,1} - x_{it+1,1}\right) b_1 + \left(x_{it,2} - x_{it+1,2}\right) b_2 + r\left(y_{it-1} - y_{it+2}\right)\right]/s\right)}\right),$$

where we impose restriction $\sqrt{b_1^2 + b_2^2} = 1$ and add the scale parameter of the logistic distribution, $s > 0$, to estimate.[21]

6.2. Tuning Parameters and Computation

There is only one tuning parameter used for estimation, namely, h_n, in the objective function (3.5). In Assumption 5, we restrict $nh_n^4 \to 0$, so that the bias term (of order h_n^2) is a small order term of $(nh_n)^{-2/3}$. Because the convergence rate of $\hat{\gamma}$ is $(nh_n)^{-1/3}$, the condition, $nh_n^4 \to 0$, makes the bias term much smaller than the convergence rate. To attain a faster convergence rate, we set h_n as large as possible, and thus set $h_n = n^{-1/4} (\log n)^{-1}$.

For the numerical bootstrap, we have one additional tuning parameter ε_n. As recommended in Hong and Li (2020), we set ε_n proportional to $n^{-2/3} \log n$ for the inferences of $\hat{\beta}$ and $\hat{\gamma}$. Apparently, ε_n of this order satisfies the additional requirements for $\hat{\gamma}^*$ that $\varepsilon_n^{-1} h_n \to \infty$ and $\varepsilon_n^{-1} h_n^4 \to 0$. To check how sensitive the

[21] HK normalize s to one, but do not require (b_1, b_2) to be on the unit circle. Both methods of scale normalization are equivalent.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

procedure is to the choice of ε_n, we conduct the procedure with $\varepsilon_n = c \cdot n^{-2/3} \log n$ and $c = 0.8, 0.9, 1.0, 1.1,$ and 1.2.

Following the recommendation in HK, we adopt the bandwidth $\sigma_n = c \cdot n^{-1/(4+k)}$ for HK's estimators, where k is the dimension of x_{it}. We conduct experiments with $c = 1, 2, 3, 4$, and report the simulation results corresponding to $c = 3$. For this value, the HK estimators of γ exhibit the smallest bias and have relatively smaller root mean squared errors among all tested values.

For all simulation designs, we employ the DE algorithm (using the `DEoptim` package in R, see Remark 3.2) to compute our proposed estimators and those of HK. To ensure efficient convergence and robustness, we set the lower and upper bounds for searching each parameter to $[-3, 3]$, the maximum number of iterations to 500, and the relative convergence tolerance to 10^{-8}. The DE algorithm does not require explicit initial values; instead, it randomly assigns $NP \times$ (the number of parameters) initial values, where we adopt the default value of $NP = 10$ in `DEoptim`. Additionally, we use the default settings for the other algorithm controls. In all simulation runs for each estimator, we consistently observe successful convergence of the algorithm. All estimators can be computed very quickly for all sample sizes considered. Using an Intel® Core™ i7-4790 processor, each replication takes only seconds to complete.

6.3. Simulation Results

We normalize the preference coefficients β on exogenous covariates to 1 in Euclidean norm. Because of this normalization, we lose one degree of freedom in the parameter space. As a result, we report only the results for (β_2, γ) in Designs 1 and 2 and the results for $(\beta_2, \ldots, \beta_k, \gamma)$ in Design k, for $k = 3, 4,$ and 5.

We report the mean bias (BIAS), the standard deviation (STD), the median absolute deviation (MAD), and the root mean squared error (RMSE) for $\hat{\beta}$ and $\hat{\gamma}$. All results are expressed as percentages of the true values of the parameters, so that the results are independent of how we normalize the parameters.[22] For inference, we report the coverage rates (COV) of the true values and lengths (LEN) of the 95% CIs for the numerical inference procedure for our method only. Furthermore, we report the computation time (TIME) for one replication of each estimator for Design 1, and omit this for other designs, owing to their similarity.

Results for Design 1 are reported in tables numbered "1," and so on for other designs. We report the performance of the estimators and the numerical bootstrap procedure in tables labeled "A" and "B," respectively. For example, Table C1A in the Appendix C reports the performance of the estimators for Design 1. The results for our estimator are denoted as "OY" in the tables (Tables C1B, C2A, and C2B). The parametric and semiparametric estimators in HK are designated as "HK1" and "HK2," respectively. Due to space limitations, only the results of Designs 1 and 2 are presented in Appendix C (Tables C1A, C1B, and C1C for Design 1

[22] We thank the co-editor for this suggestion.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

and Tables C2A and C2B for Design 2). The results of Designs 3–5, along with additional simulation studies, are reported in Appendix F of the Supplementary Material. In what follows, we briefly summarize our findings.

The RMSEs of $\hat{\beta}$ and $\hat{\gamma}$ become smaller as the sample size increases in all designs, with the RMSE of $\hat{\gamma}$ slightly greater than that of $\hat{\beta}$. This shows the consistency of our estimators, though the rates of convergence are clearly slower than $\sqrt{n}$. The numerical bootstrap inference procedures perform reasonably well in all designs. In general, they yield shrinking CIs with coverage rates approaching 95% as the sample size grows. The coverage rates of these CIs are greater than 90%, but are slightly lower than 95% in most cases. The coverage rates of the CIs for γ do not perform as well as those for β, which is not surprising, considering the complication of using two tuning parameters. The inference procedure is not very sensitive to the choice of tuning parameters.

Despite Design 2 not satisfying Assumption SI or SD, our proposed estimators still perform reasonably well in this setting.[23] Surprisingly, its performance is similar to that of Design 1, where these assumptions are completely fulfilled. This finding indicates that our method exhibits certain robustness, and can function effectively even when the two sufficient identifying assumptions are not met. Furthermore, the results from Designs 3–5 provide evidence supporting our asymptotic analysis. In these designs, we observe that the convergence rates of our estimators remain relatively stable as the dimension of the model increases.

The HK1 estimator (HK's Logit estimator) performs the best for Designs 1 and 2, which is not surprising, because the error terms are scaled logistic. Our proposed estimators exhibit higher RMSEs compared with those of HK1, approximately more than twice for these designs. For Designs 1 and 2, the HK2 estimator demonstrates finite-sample performance similar to our proposed method.[24] However, when the number of regressors increases in Design 3, our estimator outperforms HK2, particularly in estimating the parameter γ. In this setting, the RMSEs of our estimators become about 50% greater than those of HK1. In Design 4, where there are two more regressors than in Design 1, our estimators' RMSEs are comparable with those of HK1 for all sample sizes considered. Notably, for sample sizes of $n = 10,000$ and $20,000$, our RMSEs are about 40% lower than those of HK2, demonstrating the advantage of our method in high-dimensional settings. In Design 5, with three more regressors than Design 1, our RMSEs are slightly lower than those of HK1 and only half of HK2's RMSEs for sample sizes of $n = 10,000$ and $20,000$. These findings highlight the favorable properties of our method, particularly its resilience to the curse of dimensionality.

We also conduct additional simulations to assess how our estimators perform on a relatively small sample size of $n = 1,000$ for Design 1. The results presented in Table C1C in the Appendix C suggest that our estimators do not perform poorly

[23] These assumptions guarantee that the utility index $x_i'\beta$ rank orders the (conditional) probabilities in (2.2). Relaxing them could potentially invalidate the "if and only if" result in equation (2.2) and cause bias.

[24] In particular, our method exhibits smaller RMSEs for γ, and larger RMSEs for β compared to HK2.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

in terms of parameter estimation. The RMSEs are 30%-35% of the true parameter values, suggesting reasonable accuracy in estimation, despite the smaller sample size. For inference, we report only the results for $c = 1$. We can see the CIs have lower coverage of approximately 85%. In conclusion, our estimators perform reasonably well for this sample size, but the CIs may be too short.

A final note is that the serial dependence of x_{it} has limited impact on the estimation and inference, as demonstrated by the results associated with Design 2 and additional simulation studies for higher-dimensional designs presented in Appendix F of the Supplementary Material.

7. CONCLUSIONS

This paper presents new identification results for preference parameters in panel data binary choice models that allow for both fixed effects (Heckman's "spurious" state dependence) and lagged dependent variables ("true" state dependence). The same semiparametric random utility framework as in Honoré and Kyriazidou (2000) is considered. A key innovation in this paper is the assertion that, given additional restrictions on the dynamic process of observed covariates and the tail behavior of the error distribution, the point identification no longer needs element-by-element matching of regressors over time, in contrast to the method proposed in Honoré and Kyriazidou (2000). Our approach requires a minimum panel length of five ($T \geq 4$), which fits in most empirical settings. Our identification arguments motivate a two-step estimation procedure, adapting Manski's MS estimator. The proposed estimators are consistent with rates of convergence independent of the model dimension, unlike the estimator proposed in Honoré and Kyriazidou (2000). We further derive the limiting distributions of the proposed estimators, which are non-Gaussian, aligning with existing literature. We justify the use of several bootstrap procedures for conducting statistical inference. A Monte Carlo study indicates that our estimators and inference procedures perform well in finite samples.

This paper leaves some open questions for future research. For example, it might be worthwhile extending the framework in this paper to study the identification with more than one lag of the dependent variable or the identification in panel data multinomial response models.

APPENDIXES

These appendixes are organized as follows: In Appendix A, we present proofs for identification, including those for Propositions 2.1 and 2.2, along with the necessary lemmas. Additionally, we provide a roadmap for a key step in these proofs. In Appendix B, we establish the asymptotic theory of our estimators, as summarized in Theorem 4.1, and include the technical lemmas required for this proof in the same section. Appendix C compiles tables summarizing simulation results for Designs 1 and 2.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

The Supplementary Material contains Appendixes D–F. In Appendix D of the Supplementary Material, we prove all technical lemmas used in Appendixes A and B. In Appendix E of the Supplementary Material, we provide technical details for Section 5. Appendix F of the Supplementary Material presents simulation results for Designs 3–5 and supplementary simulation studies (Designs 6–8).

A. Technical Lemmas and Main Proofs for Identification

Building on the results of Lemmas A.1 and A.2, Lemma A.3 establishes the identification inequality (2.2) under Assumptions A and SD. Lemma A.4 shows that (2.2) also holds under Assumptions A and SI. We present these lemmas below and leave their proofs to Appendix D of the Supplementary Material. Based on these results, we prove Propositions 2.1 and 2.2. Throughout this appendix, we assume $\gamma < 0$. The proofs for the case with $\gamma \geq 0$ are symmetric. We omit them for conciseness.

A.1. *Roadmap for Establishing Moment Inequality (2.2)*

As indicated in the main text of the paper, the key is to establish the identifying inequality (2.2). Here, we use the simplest case with $T = 4$ to illustrate how Assumption SI, together with Assumption A, can ensure this inequality. The proof for the most general case is lengthy but uses all analogous arguments. We defer it to subsequent sections.

Our proof will repeatedly use the following identities: For any events A, B, and C,

$$P(A|B\cap C) = \frac{P(A\cap B\cap C)}{P(B\cap C)} = \frac{[P(A\cap B\cap C)/P(A\cap C)][P(A\cap C)/P(C)]}{P(B\cap C)/P(C)}$$
$$= \frac{P(B|A\cap C)P(A|C)}{P(B|C)}; \tag{A.1}$$

if event A implies event B,

$A \subseteq B$ and $A\cap B = A$;

if $A \perp B|C$,

$$P(A|B\cap C) = P(A|C); \tag{A.2}$$

if $A \subseteq B|C$,

$$P(A\cap B|C) = P(A|C); \tag{A.3}$$

and for any partition $\{B_1, \ldots, B_m\}$ of the sample space,

$$P(A|C) = \sum_{k=1}^{m} P(A\cap B_j|C) = \sum_{k=1}^{m} P(A|B_j\cap C)P(B_j|C). \tag{A.4}$$

The following result is useful:

$$(x_{t+j}, \epsilon_{t+j}) \perp y_t|\alpha,$$

holds for any $j > 0$, due to $\{x_t, \epsilon_t\} \perp \{x_s, \epsilon_s\}|\alpha$ for $s \neq t$ and the fact that y_t is a function of current and previous periods of $\{x_s, \epsilon_s\}$.

We present the proof first and defer the explanations of $\overset{(\cdot)}{=}$ to the end.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

Consider the conditional probability $P(y_1 = 1|x_1, x_3, y_0 = y_2 = y_4 = 1, \alpha)$. We can write

$$\begin{aligned}
&P(y_1 = 1|x_1, x_3, y_0 = y_2 = y_4 = 1, \alpha) \\
&\overset{\text{(i)}}{=} \frac{P(y_4 = 1|x_1, x_3, y_0 = y_1 = y_2 = 1, \alpha) P(y_1 = 1|x_1, x_3, y_0 = y_2 = 1, \alpha)}{P(y_4 = 1|x_1, x_3, y_0 = y_2 = 1, \alpha)} \\
&\overset{\text{(ii)}}{=} P(y_1 = 1|x_1, x_3, y_0 = y_2 = 1, \alpha) \\
&\overset{\text{(iii)}}{=} P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha). \qquad \textbf{(A.5)}
\end{aligned}$$

In what follows, we assume w.l.o.g. that $\gamma < 0$. The proof for the case $\gamma \geq 0$ is symmetric. We define a partition of the sample space:

$$E_{2,1} = \{\epsilon_2 < x_2'\beta + \gamma + \alpha\}, E_{2,2} = \{x_2'\beta + \gamma + \alpha \leq \epsilon_2 < x_2'\beta + \alpha\}, \text{and } E_{2,3} = \{\epsilon_2 \geq x_2'\beta + \alpha\}.$$

From the model, $E_{2,1}$ implies $\{y_2 = 1\}$, so $E_{2,1} \subseteq \{y_2 = 1\}$ and $E_{2,1} \cap \{y_2 = 1\} = E_{2,1}$. Similarly, $E_{2,3}$ implies $\{y_2 = 0\}$, so $E_{2,3} \subseteq \{y_2 = 0\}$ and $E_{2,3} \cap \{y_2 = 0\} = E_{2,3}$.

Then, we use this partition and (A.4) to write

$$\begin{aligned}
&P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha) \\
&\quad = \sum_{k=1}^{3} P(y_1 = 1|x_1, y_0 = y_2 = 1, E_{2,k}, \alpha) P(E_{2,k}|x_1, y_0 = y_2 = 1, \alpha) \\
&\quad \overset{\text{(iv)}}{=} P(y_1 = 1|x_1, y_0 = y_2 = 1, E_{2,1}, \alpha) P(E_{2,1}|x_1, y_0 = y_2 = 1, \alpha) \\
&\quad \overset{\text{(v)}}{=} P(y_1 = 1|x_1, y_0 = 1, E_{2,1}, \alpha) P(E_{2,1}|x_1, y_0 = y_2 = 1, \alpha) \\
&\quad \overset{\text{(vi)}}{=} P(y_1 = 1|x_1, y_0 = 1, \alpha) P(E_{2,1}|x_1, y_0 = y_2 = 1, \alpha) \\
&\quad \overset{\text{(vii)}}{=} F_{\epsilon|\alpha}(x_1'\beta + \gamma + \alpha) P(E_{2,1}|x_1, y_0 = y_2 = 1, \alpha). \qquad \textbf{(A.6)}
\end{aligned}$$

Applying (A.4) to the term $P(E_{2,1}|x_1, y_0 = y_2 = 1, \alpha)$ gives

$$\begin{aligned}
&P(E_{2,1}|x_1, y_0 = y_2 = 1, \alpha) \\
&= P(E_{2,1} \cap \{y_1 = 1\}|x_1, y_0 = y_2 = 1, \alpha) + P(E_{2,1} \cap \{y_1 = 0\}|x_1, y_0 = y_2 = 1, \alpha) \\
&\overset{\text{(viii)}}{=} P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha) + P(E_{2,1}|x_1, y_0 = y_2 = 1, y_1 = 0, \alpha) P(y_1 = 0|x_1, y_0 = y_2 = 1, \alpha) \\
&\overset{\text{(ix)}}{=} P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha) \\
&\quad + \frac{P(E_{2,1} \cap \{y_2 = 1\}|x_1, y_0 = 1, y_1 = 0, \alpha)}{P(y_2 = 1|x_1, y_0 = 1, y_1 = 0, \alpha)} [1 - P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha)] \\
&\overset{\text{(x)}}{=} P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha) + \frac{P(E_{2,1}|x_1, y_0 = 1, y_1 = 0, \alpha)}{P(y_2 = 1|x_1, y_0 = 1, y_1 = 0, \alpha)} [1 - P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha)] \\
&\overset{\text{(xi)}}{=} P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha) + \frac{P(E_{2,1}|\alpha)}{P(y_2 = 1|y_1 = 0, \alpha)} [1 - P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha)]. \qquad \textbf{(A.7)}
\end{aligned}$$

If we set $\Delta = P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha)$ as an unknown, (A.6) and (A.7) imply

$$\Delta = F_{\epsilon|\alpha}(x_1'\beta + \gamma + \alpha) \left[\Delta + \frac{P(E_{2,1}|\alpha)}{P(y_2 = 1|y_1 = 0, \alpha)} (1 - \Delta) \right].$$

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

Solve Δ out and apply (A.5), we obtain

$$\begin{aligned} &P(y_1 = 1|x_1, x_3, y_0 = y_2 = y_4 = 1, \alpha) \\ &= P(y_1 = 1|x_1, y_0 = y_2 = 1, \alpha) = \Delta \\ &= \frac{P(E_{2,1}|\alpha)/P(y_2 = 1|y_1 = 0, \alpha)}{1/F_{\epsilon|\alpha}(x_1'\beta + \gamma + \alpha) + 1 - P(E_{2,1}|\alpha)/P(y_2 = 1|y_1 = 0, \alpha)}. \end{aligned} \tag{A.8}$$

Furthermore, applying arguments for (A.5)–(A.7) to $P(y_3 = 1|x_1, x_3, y_0 = y_2 = y_4 = 1, \alpha)$ yields

$$\begin{aligned} &P(y_3 = 1|x_1, x_3, y_0 = y_2 = y_4 = 1, \alpha) \\ &= \frac{P(E_{4,1}|\alpha)/P(y_4 = 1|y_3 = 0, \alpha)}{1/F_{\epsilon|\alpha}(x_3'\beta + \gamma + \alpha) + 1 - P(E_{4,1}|\alpha)/P(y_4 = 1|y_3 = 0, \alpha)}, \end{aligned} \tag{A.9}$$

with $E_{4,1} = \{\epsilon_4 < x_4'\beta + \gamma + \alpha\}$. Note that Assumptions A(a) and SI(b) imply $P(E_{2,1}|\alpha)/P(y_2 = 1|y_1 = 0, \alpha) = P(E_{4,1}|\alpha)/P(y_4 = 1|y_3 = 0, \alpha)$, and Assumption A(b) guarantees the monotonicity of $F_{\epsilon|\alpha}(\cdot)$. Then identifying inequality (2.2) follows by putting all these results together and comparing (A.8) and (A.9).

To sum up, Assumption SI(a) eliminates the effects of x_s on y_t through its dependence on x_t for all $t \neq s$, and Assumption SI(b) is placed to ensure that the probabilities in (2.2) do not have time-varying representations. Using similar arguments, inequality (2.2) can be established for general cases.

We provide the explanations for $\overset{(\cdot)}{=}$ in the following.

Equality (i) in (A.5) follows from (A.1) by setting $A = \{y_1 = 1\}, B = \{y_4 = 1\}$, and $C = \{x_1, x_3, y_0 = y_2 = 1, \alpha\}$.

Equality (ii) holds due to (A.2) and $\{y_4 = 1\} \perp \{y_1 = 1\} | \{x_1, x_3, y_0 = y_2 = 1, \alpha\}$. To see why this conditional independence holds, recall that by model (2.1)

$$\begin{aligned} y_4 &= 1\left[x_4'\beta + \gamma y_3 + \alpha - \epsilon_4 \geq 0\right] \\ &= 1\left[x_4'\beta + \gamma 1\left[x_3'\beta + \gamma y_2 + \alpha - \epsilon_3 \geq 0\right] + \alpha - \epsilon_4 \geq 0\right], \end{aligned}$$

and so conditioning on $\{x_1, x_3, y_0 = y_2 = 1, \alpha\}$, the random terms remained in y_4 are $(x_4, \epsilon_3, \epsilon_4)$. Then Assumptions A(a) and SI(a), where we assume $\epsilon^T \perp \left(x^T, y_0\right)|\alpha$, and $\{x_t, \epsilon_t\} \perp \{x_s, \epsilon_s\}|\alpha$ for $s \neq t$, imply

$$(x_4, \epsilon_3, \epsilon_4) \perp \{x_1, x_3, y_0 = y_2 = 1\},\ \{y_1 = 1\}|\alpha,$$

and thus

$$(x_4, \epsilon_3, \epsilon_4) \perp \{y_1 = 1\} | \{x_1, x_3, y_0 = y_2 = 1, \alpha\},$$

implying

$$\{y_4 = 1\} \perp \{y_1 = 1\} | \{x_1, x_3, y_0 = y_2 = 1, \alpha\}.$$

Equality (iii) follows by $\{y_1 = 1\} \perp x_3 | \{x_1, y_0 = y_2 = 1, \alpha\}$ which holds for the same reason as above.

Equality (iv) is due to $P(E_{2,3}|x_1, y_0 = y_2 = 1, \alpha) = 0$ (since $E_{2,3}$ implies $\{y_2 = 0\}$ and thus $E_{2,3} \subseteq \{y_2 = 0\}$), and $P(y_1 = 1|x_1, y_0 = y_2 = 1, E_{2,2}, \alpha) = 0$ (since $\{y_2 = 1\} \cap E_{2,2}$ implies $\{y_1 = 0\}$) .

Equality (v) follows by $E_{2,1} \subseteq \{y_2 = 1\}$.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

Equality (vi) holds because $P(y_1 = 1|x_1, y_0 = 1, E_{2,1}, \alpha) = P(y_1 = 1|x_1, y_0 = 1, \alpha)$ implied by (A.2) (letting $A = \{y_1 = 1\}$, $B = E_{2,1}$, and $C = \{x_1, y_0 = 1, \alpha\}$).

Equality (vii) holds because of Assumption A(a) that $\epsilon^T \perp \left(x^T, y_0\right)|\alpha$.

Equality (viii) holds because conditional on $\{y_2 = 1\}$ and $\gamma < 0$, $\{y_1 = 1\}$ implies $E_{2,1}$, and thus $\{y_1 = 1\} \subseteq E_{2,1}$ conditional on $\{x_1, y_0 = y_2 = 1, \alpha\}$.

Equality (ix) holds due to fact that $P(A|B \cap C) = P(A \cap B|C)/P(B|C)$.

Equality (x) holds since $E_{2,1}$ implies $\{y_2 = 1\}$ and thus $E_{2,1} \subseteq \{y_2 = 1\}$.

Equality (xi) follows by model (2.1), Assumption A(a), Assumption SI(a), and applying (A.2) (letting $A = E_{2,1}$, $B = \{x_1, y_0 = 1, y_1 = 0\}$, and $C = \{\alpha\}$ for $P(E_{2,1}|x_1, y_0 = 1, y_1 = 0, \alpha)$, and $A = \{y_2 = 1\}$, $B = \{x_1, y_0 = 1\}$, and $C = \{y_1 = 0, \alpha\}$ for $P(y_2 = 1|x_1, y_0 = 1, y_1 = 0, \alpha)$).

A.2. *Technical Lemmas and Main Proofs for Identification*

Now we rigorously prove our identification results (Propositions 2.1 and 2.2). Before that, we first list necessary technical lemmas whose proofs are presented in Appendix D of the Supplementary Material. For each $t \in \mathcal{T}$, define the following partition of the sample space:[25]

$$E_{t,1} = \{\epsilon_t < w_t + \gamma + \alpha\}, E_{t,2} = \{w_t + \gamma + \alpha \leq \epsilon_t < w_t + \alpha\}, E_{t,3} = \{\epsilon_t \geq w_t + \alpha\}.$$

LEMMA A.1. *Let $s, t \in \mathcal{T}$ such that $t \geq s+2$. Under Assumption A, the following equalities hold for both $\tau = s$ and $\tau = t$:*

$$\begin{aligned} &P(y_\tau = 1|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 1, \alpha) \\ &= F_{\epsilon|\alpha}(w_\tau + \gamma y_{\tau-1} + \alpha) P(E_{\tau+1,1}|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 1, \alpha), \end{aligned} \tag{A.10}$$

and

$$\begin{aligned} &P(y_\tau = 1|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 0, \alpha) \\ &= P(E_{\tau+1,2}|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 0, \alpha) \\ &\quad + F_{\epsilon|\alpha}(w_\tau + \gamma y_{\tau-1} + \alpha) P(E_{\tau+1,3}|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 0, \alpha). \end{aligned} \tag{A.11}$$

LEMMA A.2. *Let $s, t \in \mathcal{T}$ such that $t \geq s+2$. Under Assumption A, the following equalities hold for both $\tau = s$ and $\tau = t$:*

$$\begin{aligned} &P(E_{\tau+1,1}|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 1, \alpha) \\ &= P(y_\tau = 1|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 1, \alpha) \\ &\quad + \frac{F_{\epsilon|\alpha}(w_{\tau+1} + \gamma + \alpha)}{F_{\epsilon|\alpha}(w_{\tau+1} + \alpha)} [1 - P(y_\tau = 1|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 1, \alpha)], \end{aligned} \tag{A.12}$$

$$\begin{aligned} &P(E_{\tau+1,2}|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 0, \alpha) \\ &= \frac{F_{\epsilon|\alpha}(w_{s+1} + \alpha) - F_{\epsilon|\alpha}(w_{s+1} + \gamma + \alpha)}{1 - F_{\epsilon|\alpha}(w_{s+1} + \gamma + \alpha)} P(y_\tau = 1|w^T, y_{s-1} = y_{t-1}, y_{s+1} = y_{t+1} = 0, \alpha), \end{aligned} \tag{A.13}$$

[25] For the case with $\gamma \geq 0$, the proofs of Lemmas A.1–A.3 work through with the partition $E_{t,1} = \{\epsilon_t < w_t + \alpha\}, E_{t,2} = \{w_t + \alpha \leq \epsilon_t < w_t + \gamma + \alpha\}$, and $E_{t,3} = \{\epsilon_t \geq w_t + \gamma + \alpha\}$.

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

and

$$\begin{aligned}&P(E_{\tau+1,3}|w^T, y_{s-1}=y_{t-1}, y_{s+1}=y_{t+1}=0,\alpha)\\ =&\frac{1-F_{\epsilon|\alpha}(w_{\tau+1}+\alpha)}{1-F_{\epsilon|\alpha}(w_{\tau+1}+\gamma+\alpha)}P(y_\tau=1|w^T, y_{s-1}=y_{t-1}, y_{s+1}=y_{t+1}=0,\alpha)\\ &+1-P(y_\tau=1|w^T, y_{s-1}=y_{t-1}, y_{s+1}=y_{t+1}=0,\alpha). \qquad \textbf{(A.14)}\end{aligned}$$

LEMMA A.3. *If Assumptions A and SD hold, then for all* $s,t\in\mathcal{T}$,

$$P(y_t=1|w_s,w_t,y_{s-1}=y_{t-1},y_{s+1}=y_{t+1},\alpha)\geq P(y_s=1|w_s,w_t,y_{s-1}=y_{t-1},y_{s+1}=y_{t+1},\alpha)$$

if and only if $w_t\geq w_s$.

LEMMA A.4. *If Assumptions A and SI hold, then for all* $s,t\in\mathcal{T}$,

$$P(y_t=1|w_s,w_t,y_{s-1}=y_{t-1},y_{s+1}=y_{t+1},\alpha)\geq P(y_s=1|w_s,w_t,y_{s-1}=y_{t-1},y_{s+1}=y_{t+1},\alpha)$$

if and only if $w_t\geq w_s$.

We next prove Propositions 2.1 and 2.2 in order.

Proof of Proposition 2.1. The monotonic relation established in either Lemma A.3 or Lemma A.4 implies that β maximizes $Q_1(\cdot;\alpha)$. The remaining task is to show the uniqueness of β in $\mathcal{B}$, that is, $Q_1(b;\alpha)=Q_1(\beta;\alpha)$ implies $b=\beta$. Here, we assume $\beta_1>0$ w.l.o.g. as the case $\beta_1<0$ is symmetric.

First, note that for any $b\in\mathcal{B}$ such that $Q_1(b;\alpha)=Q_1(\beta;\alpha)$, if

$$P([x_{ts,1}b_1+\tilde{x}_{ts}'\tilde{b}<0<x_{ts,1}\beta_1+\tilde{x}_{ts}'\tilde{\beta}]\cup[x_{ts,1}\beta_1+\tilde{x}_{ts}'\tilde{\beta}<0<x_{ts,1}b_1+\tilde{x}_{ts}'\tilde{b}])>0,$$

then β and b will yield different realized values of the sign function in $Q_1(\cdot;\alpha)$ with strictly positive probability, and thus $Q_1(\beta;\alpha)>Q_1(b;\alpha)$. It then follows that $b_1>0$ must hold, for otherwise by Assumption A(c), we have

$$P(x_{ts,1}b_1+\tilde{x}_{ts}'\tilde{b}<0<x_{ts,1}\beta_1+\tilde{x}_{ts}'\tilde{\beta})=P(x_{ts,1}>-\tilde{x}_{ts}'\tilde{b}/b_1, x_{ts,1}>-\tilde{x}_{ts}'\tilde{\beta}/\beta_1)>0.$$

Then focusing on the case with $b_1>0$, we can write

$$\begin{aligned}&P([x_{ts,1}b_1+\tilde{x}_{ts}'\tilde{b}<0<x_{ts,1}\beta_1+\tilde{x}_{ts}'\tilde{\beta}]\cup[x_{ts,1}\beta_1+\tilde{x}_{ts}'\tilde{\beta}<0<x_{ts,1}b_1+\tilde{x}_{ts}'\tilde{b}])\\ =&P([-\tilde{x}_{ts}'\tilde{\beta}/\beta_1<x_{ts,1}<-\tilde{x}_{ts}'\tilde{b}/b_1]\cup[-\tilde{x}_{ts}'\tilde{b}/b_1<x_{ts,1}<-\tilde{x}_{ts}'\tilde{\beta}/\beta_1]),\end{aligned}$$

which implies that to make $Q_1(b;\alpha)=Q_1(\beta;\alpha)$ hold we must have $P(\tilde{x}_{ts}'\tilde{\beta}/\beta_1=\tilde{x}_{ts}'\tilde{b}/b_1)=1$ by Assumption A(c).

However, whenever b is not a scalar multiple of β, $P(\tilde{x}_{ts}'\tilde{\beta}/\beta_1=\tilde{x}_{ts}'\tilde{b}/b_1)=1$ implies that $\tilde{\mathcal{X}}_{ts}$ is contained in a proper linear subspace of $\mathbb{R}^{K-1}$ almost everywhere, violating Assumption A(d). As a result, we must have b as a scalar multiple of β, which leads to the desired result $b=\beta$ as $\|b\|_2=\|\beta\|_2=1$ by the construction of the parameter space $\mathcal{B}$ in Assumption A(e). □

Proof of Proposition 2.2. The proof uses the insight of HK. Here, we only prove case (ii) of Proposition 2.2 for $t>s+1$ as the same method can be applied to case (i) where s and t are adjacent. Note that it also suffices to prove that γ uniquely maximizes the following

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

population objective function conditional on α:

$$Q_{2,2}(\gamma;\beta,\alpha) \equiv \mathbb{E}\{[P(A|x^T, w_{s+1}=w_{t+1}, y_{s+1}=y_{t+1}, \alpha) - P(B|x^T, w_{s+1}=w_{t+1}, y_{s+1}=y_{t+1}, \alpha)] \times sgn((w_t - w_s) + r(d_{t-1} - d_{s-1}))|\alpha\}.$$

First, note that under Assumptions A(a) and A(b), we can write

$$\begin{aligned}
&P(A|x^T, w_{s+1}=w_{t+1}=w, y_{s+1}=y_{t+1}=d, \alpha) \\
&= p_0(x^T,\alpha)^{d_0}(1-p_0(x^T,\alpha))^{1-d_0} \times F_{\epsilon|\alpha}(w_1+\gamma d_0+\alpha)^{d_1}(1-F_{\epsilon|\alpha}(w_1+\gamma d_0+\alpha))^{1-d_1} \\
&\quad \times \cdots \times (1-F_{\epsilon|\alpha}(w_s+\gamma d_{s-1}+\alpha)) \times F_{\epsilon|\alpha}(w+\alpha)^{d}(1-F_{\epsilon|\alpha}(w+\alpha))^{1-d} \\
&\quad \times \cdots \times F_{\epsilon|\alpha}(w_t+\gamma d_{t-1}+\alpha) \times F_{\epsilon|\alpha}(w+\gamma+\alpha)^{d}(1-F_{\epsilon|\alpha}(w+\gamma+\alpha))^{1-d} \\
&\quad \times \cdots \times F_{\epsilon|\alpha}(w_T+\gamma d_{T-1}+\alpha)^{d_T}(1-F_{\epsilon|\alpha}(w_T+\gamma d_{T-1}+\alpha))^{1-d_T}
\end{aligned}$$

for all $w \in \mathbb{R}$ and $d \in \{0,1\}$, and similarly,

$$\begin{aligned}
&P(B|x^T, w_{s+1}=w_{t+1}=w, y_{s+1}=y_{t+1}=d, \alpha) \\
&= p_0(x^T,\alpha)^{d_0}(1-p_0(x^T,\alpha))^{1-d_0} \times F_{\epsilon|\alpha}(w_1+\gamma d_0+\alpha)^{d_1}(1-F_{\epsilon|\alpha}(w_1+\gamma d_0+\alpha))^{1-d_1} \\
&\quad \times \cdots \times F_{\epsilon|\alpha}(w_s+\gamma d_{s-1}+\alpha) \times F_{\epsilon|\alpha}(w+\gamma+\alpha)^{d}(1-F_{\epsilon|\alpha}(w+\gamma+\alpha))^{1-d} \\
&\quad \times \cdots \times (1-F_{\epsilon|\alpha}(w_t+\gamma d_{t-1}+\alpha)) \times F_{\epsilon|\alpha}(w+\alpha)^{d}(1-F_{\epsilon|\alpha}(w+\alpha))^{1-d} \\
&\quad \times \cdots \times F_{\epsilon|\alpha}(w_T+\gamma d_{T-1}+\alpha)^{d_T}(1-F_{\epsilon|\alpha}(w_T+\gamma d_{T-1}+\alpha))^{1-d_T}.
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
&\frac{P(A|x^T, w_{s+1}=w_{t+1}=w, y_{s+1}=y_{t+1}=d, \alpha)}{P(B|x^T, w_{s+1}=w_{t+1}=w, y_{s+1}=y_{t+1}=d, \alpha)} \\
&= \frac{(1-F_{\epsilon|\alpha}(w_s+\gamma d_{s-1}+\alpha)) \times F_{\epsilon|\alpha}(w_t+\gamma d_{t-1}+\alpha)}{F_{\epsilon|\alpha}(w_s+\gamma d_{s-1}+\alpha) \times (1-F_{\epsilon|\alpha}(w_t+\gamma d_{t-1}+\alpha))} \\
&\quad \times \frac{F_{\epsilon|\alpha}(w+\alpha)^{d}(1-F_{\epsilon|\alpha}(w+\alpha))^{1-d} \times F_{\epsilon|\alpha}(w+\gamma+\alpha)^{d}(1-F_{\epsilon|\alpha}(w+\gamma+\alpha))^{1-d}}{F_{\epsilon|\alpha}(w+\gamma+\alpha)^{d}(1-F_{\epsilon|\alpha}(w+\gamma+\alpha))^{1-d} \times F_{\epsilon|\alpha}(w+\alpha)^{d}(1-F_{\epsilon|\alpha}(w+\alpha))^{1-d}} \\
&= \frac{(1-F_{\epsilon|\alpha}(w_s+\gamma d_{s-1}+\alpha)) \times F_{\epsilon|\alpha}(w_t+\gamma d_{t-1}+\alpha)}{F_{\epsilon|\alpha}(w_s+\gamma d_{s-1}+\alpha) \times (1-F_{\epsilon|\alpha}(w_t+\gamma d_{t-1}+\alpha))}
\end{aligned}$$

and therefore,

$$P(A|x^T, w_{s+1}=w_{t+1}=w, y_{s+1}=y_{t+1}=d, \alpha) \geq P(B|x^T, w_{s+1}=w_{t+1}=w, y_{s+1}=y_{t+1}=d, \alpha)$$

if and only if $w_t + \gamma d_{t-1} \geq w_s + \gamma d_{s-1}$, which implies that γ maximizes $Q_{2,2}(\gamma;\beta,\alpha)$.

The remaining task is to show that γ is unique in $\mathcal{R}$. Suppose that there exists an $r \in \mathcal{R} \setminus \{\gamma\}$ such that $Q_{2,2}(r;\beta,\alpha) = Q_{2,2}(\gamma;\beta,\alpha)$. Note that the value of r (and γ) affects $Q_{2,2}(\cdot;\beta,\alpha)$ only when $d_{s-1} \neq d_{t-1}$. Here, we assume that $d_{t-1}=1$ and $d_{s-1}=0$ (the case with $d_{t-1}=0$ and $d_{s-1}=1$ is symmetric). Then by Assumption A(c), the following probability is nonzero:

$$P([-\gamma < w_t - w_s < -r] \cup [-r < w_t - w_s < -\gamma]).$$

Consequently, γ and r yield different realized values of the sign function in objective function $Q_{2,2}(\cdot;\beta,\alpha)$ with strictly positive probability, and hence $Q_{2,2}(r;\beta,\alpha) < Q_{2,2}(\gamma;\beta,\alpha)$,

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

a contradiction. Then we can conclude that $Q_{2,2}(r;\beta,\alpha) = Q_{2,2}(\gamma;\beta,\alpha)$ if and only if $r=\gamma$, or equivalently γ uniquely maximizes $Q_{2,2}(\cdot;\beta,\alpha)$ in $\mathcal{R}$. □

B. Technical Lemmas and Main Proofs for Asymptotics

In this section, we define a few technical terms and a few more technical notations, present some technical lemmas, and prove our main asymptotic theory, Theorem 4.1. The proofs for the technical lemmas are relegated to Appendix D of the Supplementary Material.

The outline of the proof of Theorem 4.1 is as follows. Lemmas B.1 and B.2 verify the technical conditions as required in Seo and Otsu (2018). Those conditions can ensure the class of functions is *manageable* as in Kim and Pollard (1990). After that, the maximal inequalities and asymptotics in Seo and Otsu (2018) can be readily applied to our estimator. Lemma B.4 deals with the impact of using $\hat{\beta}$ on estimating $\hat{\gamma}$, using maximal inequalities established in Seo and Otsu (2018). Lemmas B.3 and B.5 obtain the technical terms for the final asymptotics for $\hat{\beta}$ and $\hat{\gamma}$.

Let c and C denote some constants that may vary from line to line. $\mathbb{E}_n$ denotes the expectation conditional on observations being fixed. $\rightsquigarrow$ denotes weakly convergence in the sense of van der Vaart and Wellner (1996). Let

$$\mathbb{G}_n(f_{ni}) \equiv n^{1/2}\sum_{i=1}^{n}[f_{ni}-\mathbb{E}_n(f_n)],$$

for any f_{ni}. To facilitate calculation, occasionally we may decompose covariate x into $\varpi\beta + x_\beta$ with a scalar ϖ and x_β orthogonal to β.

We define the following technical terms used in lemmas:

$$Z_{n,1}(\boldsymbol{s}) \equiv n^{2/3}\cdot n^{-1}\sum_{i=1}^{n}\xi_i\left(\beta + \boldsymbol{s}n^{-1/3}\right),$$

$$Z_{n,2}(s) \equiv (nh_n)^{2/3}\cdot n^{-1}\sum_{i=1}^{n}\varsigma_{ni}\left(\gamma + s(nh_n)^{-1/3},\beta\right),$$

and

$$\hat{Z}_{n,2}(s) \equiv (nh_n)^{2/3}\cdot n^{-1}\sum_{i=1}^{n}\varsigma_{ni}\left(\gamma + s(nh_n)^{-1/3},\hat{\beta}\right).$$

Note that the $\boldsymbol{s}$ in $Z_{n,1}(\boldsymbol{s})$ is a $K\times 1$ vector, and the s in $Z_{n,2}(s)$ and $\hat{Z}_{n,2}(s)$ is a scalar.

LEMMA B.1. *Suppose Assumptions A, SI (or SD), and 3 hold. Then $\xi_i(b)$ satisfies Assumption M in Seo and Otsu (2018).*

LEMMA B.2. *Suppose Assumptions A, SI (or SD) and 3–5 hold. Then $\varsigma_{ni}(r,b)$ satisfies Assumption M in Seo and Otsu (2018).*

LEMMA B.3. *Suppose Assumptions A and 3 hold. Then*

$$\lim_{n\to\infty} n^{2/3}\mathbb{E}\left(\xi_i\left(\beta+\boldsymbol{s}n^{-1/3}\right)\right) = \frac{1}{2}\boldsymbol{s}'V_1\boldsymbol{s},$$

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

and

$$\lim_{n\to\infty} n^{1/3}\mathbb{E}\left[\xi_i\left(\beta+\boldsymbol{s}n^{-1/3}\right)\xi_i\left(\beta+\boldsymbol{t}n^{-1/3}\right)\right]=H_1(\boldsymbol{s},\boldsymbol{t}).$$

V_1 is defined as

$$V_1=-\int 1\left[x_{31}'\beta=0\right]\left(\frac{\partial\kappa(x_{31})'}{\partial x_{31}}\beta\right)f_{x_{31}}(x_{31})x_{31}x_{31}'d\sigma_0, \tag{B.1}$$

with σ_0 being the surface measure on $\{x_{31}:x_{31}'\beta=0\}$ and

$$\kappa(x)=\mathbb{E}\{P(y_{i0}=y_{i2}=y_{i4}|x_{i1},x_{i3})\{\mathbb{E}[y_{i3}|y_{i2}=y_{i4},x_{i3}]-\mathbb{E}[y_{i1}|y_{i0}=y_{i2},x_{i1}]\}|x_{i31}=x\}.$$

$H_1(\boldsymbol{s},\boldsymbol{t})$ is defined as

$$H_1(\boldsymbol{s},\boldsymbol{t})=\frac{1}{2}\int_{\mathbb{R}^{K-1}}\psi(x_\beta)\left[\left|x_\beta'\boldsymbol{s}\right|+\left|x_\beta'\boldsymbol{t}\right|-\left|x_\beta'(\boldsymbol{s}-\boldsymbol{t})\right|\right]f_{x_{31}}(x_\beta)\,dx_\beta, \tag{B.2}$$

where $\boldsymbol{s},\boldsymbol{t}$ are $K\times 1$ vectors,

$$\psi(x)=\mathbb{E}\{P(y_{i0}=y_{i2}=y_{i4}|x_{i1},x_{i3})\,|\mathbb{E}[y_{i3}|y_{i2}=y_{i4},x_{i3}]-\mathbb{E}[y_{i1}|y_{i0}=y_{i2},x_{i1}]|\;|x_{i31}=x\},$$

and x_β is orthogonal to β.

LEMMA B.4. *Suppose Assumptions A and 3–5 hold. Then*

$$\hat{Z}_{n,2}(s)-Z_{n,2}(s)=o_P(1),$$

where the small order term holds uniformly over $|s|\le C$ for any positive C.

LEMMA B.5. *Suppose Assumptions A and 3–5 hold. Then*

$$\lim_{n\to\infty}(nh_n)^{2/3}\mathbb{E}\left(\varsigma_{ni}\left(\gamma+s(nh_n)^{-1/3},\beta\right)\right)=\frac{1}{2}V_2s^2,$$

and

$$\lim_{n\to\infty}(nh_n)^{1/3}\mathbb{E}\left(h_n\varsigma_{ni}\left(\gamma+s(nh_n)^{-1/3},\beta\right)\varsigma_{ni}\left(\gamma+t(nh_n)^{-1/3},\beta\right)\right)=H_2(s,t).$$

V_2 is defined as

$$\begin{aligned}V_2=&-\int_{\mathbb{R}^{K-1}}\int 1\left[x_{21}'\beta+\gamma y_{30}=0\right]\left(\frac{\partial\mathbb{E}\left(y_{21}|x_{21},y_{30},x_{32}=x_\beta\right)'}{\partial\left(y_{30},x_{21}'\right)'}\begin{pmatrix}\gamma\\ \beta\end{pmatrix}\right)\\ &f\left(x_{21},y_{30}|x_{32}=x_\beta\right)|y_{30}|\,d\sigma_0 f_{x_{32}}(x_\beta)\,dx_\beta\\ &-\int_{\mathbb{R}^{K-1}}\int 1\left[x_{32}'\beta+\gamma y_{41}=0\right]\left(\frac{\partial\mathbb{E}\left(y_{32}|x_{32},y_{41},x_{43}=x_\beta\right)'}{\partial\left(y_{41},x_{32}'\right)'}\begin{pmatrix}\gamma\\ \beta\end{pmatrix}\right)\\ &f\left(x_{32},y_{41}|x_{43}=x_\beta\right)|y_{41}|\,d\sigma_0 f_{x_{43}}(x_\beta)\,dx_\beta\end{aligned} \tag{B.3}$$

with σ_0 denoting the surface measure of $\left\{\left(x_{21}',y_{30}\right)'\middle|x_{21}'\beta+\gamma y_{30}=0\right\}$ in the first integral and the surface measure of $\left\{\left(x_{32}',y_{41}\right)'\middle|x_{32}'\beta+\gamma y_{41}=0\right\}$ in the second integral. $H_2(s,t)$ is defined as

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

$$H_2(s,t) \tag{B.4}$$
$$= \frac{1}{2}(|s|+|t|-|s-t|)\bar{\mathcal{K}}_2 \int_{\mathbb{R}^{K-1}} \{\mathbb{E}\left[|y_{21}||x_{21}'\beta = -\gamma, y_{30} = 1, x_{32} = x_\beta\right]$$
$$f\left(y_{30} = 1, x_{21}'\beta = -\gamma | x_{32} = x_\beta\right)$$
$$+\mathbb{E}\left[|y_{21}||x_{21}'\beta = \gamma, y_{30} = -1, x_{32} = x_\beta\right] f\left(y_{30} = -1, x_{21}'\beta = \gamma | x_{32} = x_\beta\right)\} f_{x_{32}}(x_\beta)\, dx_\beta$$
$$+\frac{1}{2}(|s|+|t|-|s-t|)\bar{\mathcal{K}}_2 \int_{\mathbb{R}^{K-1}} \{\mathbb{E}\left[|y_{32}||x_{32}'\beta = -\gamma, y_{41} = 1, x_{43} = x_\beta\right]$$
$$f\left(y_{41} = 1, x_{32}'\beta = -\gamma | x_{43} = x_\beta\right)$$
$$+\mathbb{E}\left[|y_{32}||x_{32}'\beta = \gamma, y_{41} = -1, x_{43} = x_\beta\right] f\left(y_{41} = -1, x_{32}'\beta = \gamma | x_{43} = x_\beta\right)\} f_{x_{43}}(x_\beta)\, dx_\beta,$$

where s,t are scalars, $\bar{\mathcal{K}}_2 = \int_{\mathbb{R}} \mathcal{K}(u)^2\, du$, and x_β is orthogonal to β.

Note that y_{30} can only take values $-1, 0,$ or 1. We assume that y_{30} takes hypothetical values in $(-1-\varepsilon, -1+\varepsilon)$, $(-\varepsilon, \varepsilon)$, and $(1-\varepsilon, 1+\varepsilon)$ for a small $\varepsilon > 0$ when calculating derivatives with respect to y_{30}. For example, $\frac{d\Phi(y_{30})}{dy_{30}}\Big|_{y_{30}=1} = \Phi'(y_{30})\big|_{y_{30}=1} = \Phi'(1)$ for any continuous differentiable Φ.

Proof of Theorem 4.1. We prove the first part of this theorem first.

Lemma B.1 verifies the key technical conditions needed for applying Theorem 1 in Seo and Otsu (2018). $\hat{\beta} - \beta = O_P\left(n^{-1/3}\right)$ by Assumption 2 and Lemma 1 in Seo and Otsu (2018).

Notice that $\hat{\beta}$ can be equivalently obtained from

$$\arg\max_{b \in \mathcal{B}} n^{2/3} \cdot n^{-1} \sum_{i=1}^{n} \xi_i \left(\beta + n^{-1/3} \cdot n^{1/3}(b-\beta)\right).$$

Intuitively, we get the asymptotics of $n^{1/3}\left(\hat{\beta} - \beta\right)$ if we can get the asymptotics of

$$Z_{n,1}(s) = n^{2/3} \cdot n^{-1} \sum_{i=1}^{n} \xi_i \left(\beta + s n^{-1/3}\right).$$

Lemma B.3 calculates the mean and covariance kernel of $Z_{n,1}(s).\xi_i(b)$ is uniformly bounded, so the Lindeberg condition for $Z_{n,1}(s)$ is satisfied. Therefore, $Z_{n,1}(s)$ is pointwise asymptotically normal. With Assumption 2, Theorem 1 in Seo and Otsu (2018) implies the equicontinuity of $Z_{n,1}(s)$, and it yields $Z_{n,1}(s) \rightsquigarrow Z_1(s)$, where $Z_1(s)$ is a Gaussian Process with continuous sample paths, expected value $-\frac{1}{2}s'V_1 s$, and covariance kernel $H_1(s,t)$ that can be calculated as in equation (B.2). As a result,

$$n^{1/3}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \arg\max_{s \in \mathbb{R}^K} Z_1(s),$$

by applying Theorem 1 in Seo and Otsu (2018).

We now prove the second part. The calculation of equation (D.31) in the proof of Lemma B.5 shows,

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

$$\mathbb{E}_n\left(\varsigma_{ni}\left(r,\hat{\beta}\right)-\varsigma_{ni}\left(\gamma,\beta\right)\right)=\frac{1}{2}\left(r-\gamma,\left(\hat{\beta}-\beta\right)'\right)\tilde{V}_2\begin{pmatrix} r-\gamma \\ \hat{\beta}-\beta \end{pmatrix} \tag{B.5}$$
$$+o\left(\left\|\left(r-\gamma,\left(\hat{\beta}-\beta\right)'\right)'\right\|_2\right)+o\left((nh_n)^{-2/3}\right),$$

where $\tilde{V}_2$ is defined in equation (D.30).

The convergence rate of $\hat{\gamma}$ is $(nh_n)^{-1/3}$, which can be seen from

$$\begin{aligned}
o_P\left((nh_n)^{-2/3}\right) &\leq n^{-1}\sum_{i=1}^{n}\varsigma_{ni}\left(\hat{\gamma},\hat{\beta}\right)-n^{-1}\sum_{i=1}^{n}\varsigma_{ni}\left(\gamma,\hat{\beta}\right)\\
&=n^{-1}\sum_{i=1}^{n}\varsigma_{ni}\left(\hat{\gamma},\hat{\beta}\right)-n^{-1}\sum_{i=1}^{n}\varsigma_{ni}\left(\gamma,\beta\right)+n^{-1}\sum_{i=1}^{n}\varsigma_{ni}\left(\gamma,\beta\right)-n^{-1}\sum_{i=1}^{n}\varsigma_{ni}\left(\gamma,\hat{\beta}\right)\\
&\leq \mathbb{E}_n\left(\varsigma_{ni}\left(\hat{\gamma},\hat{\beta}\right)-\varsigma_{ni}\left(\gamma,\beta\right)\right)+\varepsilon\left(\left(\hat{\gamma}-\gamma\right)^2+\left\|\hat{\beta}-\beta\right\|_2^2\right)+O_P\left((nh_n)^{-2/3}\right)\\
&+\mathbb{E}_n\left(\varsigma_{ni}\left(\gamma,\hat{\beta}\right)-\varsigma_{ni}\left(\gamma,\beta\right)\right)+\varepsilon\left\|\hat{\beta}-\beta\right\|_2^2+O_P\left((nh_n)^{-2/3}\right)\\
&\leq(-c+\varepsilon)\left(\left(\hat{\gamma}-\gamma\right)^2+2\left\|\hat{\beta}-\beta\right\|_2^2\right)+o\left(\left(\hat{\gamma}-\gamma\right)^2+\left\|\hat{\beta}-\beta\right\|_2^2\right)+O_P\left((nh_n)^{-2/3}\right),
\end{aligned}$$

for each $\varepsilon>0$, where the first line holds by Assumption 2, the third to fourth lines hold by applying Lemma 1 in Seo and Otsu (2018), the fifth line holds by equation (B.5). By noting $\left\|\hat{\beta}-\beta\right\|_2=O_P\left(n^{-1/3}\right)=o_P\left((nh_n)^{-1/3}\right)$, the inequality above implies

$$0\leq(-c+\varepsilon)\left(\hat{\gamma}-\gamma\right)^2+o\left(\left(\hat{\gamma}-\gamma\right)^2\right)+O_P\left((nh_n)^{-2/3}\right).$$

Taking an ε to satisfy $\varepsilon<<c$ yields that $\hat{\gamma}-\gamma=O_P\left((nh_n)^{-1/3}\right)$. So we only need to get the limiting distribution of $\hat{Z}_{n,2}(s)$.

The analysis of $\hat{Z}_{n,2}(s)$ is complicated by including the first-stage estimator $\hat{\beta}$. Lemma B.4 shows that $\hat{\beta}$ has no impact on the asymptotics of $\hat{\gamma}$. More specifically,

$$\begin{aligned}
\hat{Z}_{n,2}(s)&=Z_{n,2}(s)+o_P(1)\\
&=n^{1/6}h_n^{2/3}\mathbb{G}_n\left(\varsigma_{ni}\left(\gamma+s(nh_n)^{-1/3},\beta\right)\right)+(nh_n)^{2/3}\mathbb{E}\left(\varsigma_{ni}\left(\gamma+s(nh_n)^{-1/3},\beta\right)\right)+o_P(1),
\end{aligned} \tag{B.6}$$

where $\mathbb{G}_n\left(\varsigma_{ni}(r,b)\right)=n^{-1/2}\sum_{i=1}^{n}\left(\varsigma_{ni}(r,b)-\mathbb{E}_n\left(\varsigma_{ni}(r,b)\right)\right)$. As a result, the asymptotics is established if the weak convergence of the leading term in equation (B.6) is proved.

Lemma B.5 calculates the the mean of $(nh_n)^{2/3}\mathbb{E}\left(\varsigma_{ni}\left(\gamma+s(nh_n)^{-1/3},\beta\right)\right)$ and covariance kernel $n^{1/6}h_n^{2/3}\mathbb{G}_n\left(\varsigma_{ni}\left(\gamma+s(nh_n)^{-1/3},\beta\right)\right)$.

Note

$$\begin{aligned}
&\sum_{i=1}^{n}\left((nh_n)^{2/3}\cdot n^{-1}\right)^{2+\delta}\mathbb{E}\left[\left|\varsigma_{ni}\left(\gamma+s(nh_n)^{-1/3},\beta\right)\right|^{2+\delta}\right]\\
&=(nh_n)^{-\delta/3}\cdot(nh_n)^{1/3}\mathbb{E}\left[h_n^{1+\delta}\left|\varsigma_{ni}\left(\gamma+s(nh_n)^{-1/3},\beta\right)\right|^{2+\delta}\right]\to 0
\end{aligned}$$

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

for a small $\delta > 0$, because $(nh_n)^{-\delta/3} \to 0$ and $(nh_n)^{1/3}\mathbb{E}\left[h_n^{1+\delta}\left|\varsigma_{ni}\left(\gamma + s(nh_n)^{-1/3}, \beta\right)\right|^{2+\delta}\right] \to c$ for a finite c. This verifies the Lyapunov condition for $n^{1/6}h_n^{2/3}\mathbb{G}_n\left(\varsigma_{ni}\left(\gamma + s\left(nh_n\right)^{-1/3}, \beta\right)\right)$. Therefore, it converges to normal in distribution for each s. Lemma B.2 verifies the key technical conditions for applying Theorem 1 in Seo and Otsu (2018) to $Z_{n,2}(s)$. Together with Assumption 2, all technical conditions in Theorem 1 of Seo and Otsu (2018) are satisfied for $Z_{n,2}(s)$. That implies the stochastic equicontinuity of $Z_{n,2}(s)$ in s and

$$Z_{n,2}(s) \rightsquigarrow Z_2(s),$$

where $Z_2(s)$ is a Gaussian process with continuous path, expected value $\frac{1}{2}V_2 s^2$ and covariance kernel $H_2(s,t)$. Then, the following result follows by the continuous mapping theorem:

$$(nh_n)^{1/3}\left(\hat{\gamma} - \gamma\right) \xrightarrow{d} \arg\max_{s\in\mathbb{R}} Z_2(s). \qquad \square$$

C. Simulation Results of Designs 1 and 2

TABLE C1A. Design 1, performance of $\hat{\beta}$ and $\hat{\gamma}$.

	$n = 2,500$		$n = 5,000$		$n = 10,000$		$n = 20,000$	
	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$
OY BIAS	1.8%	−0.9%	1.7%	−0.3%	0.8%	1.5%	0.7%	0.5%
STD	20.2%	25.0%	14.7%	18.9%	11.4%	14.7%	9.1%	12.0%
MAD	15.9%	19.4%	11.7%	15.1%	9.0%	11.8%	7.3%	9.7%
RMSE	20.2%	25.0%	14.8%	18.9%	11.4%	14.8%	9.1%	12.0%
TIME	0.33 seconds		0.62 seconds		1.16 seconds		2.50 seconds	
HK1 BIAS	−0.4%	2.4%	−0.0%	1.8%	−0.2%	1.4%	−0.1%	1.2%
STD	5.7%	15.5%	4.6%	11.6%	3.6%	8.8%	2.8%	6.8%
MAD	4.6%	12.3%	3.7%	9.2%	2.9%	7.2%	2.2%	5.5%
RMSE	5.7%	15.6%	4.6%	11.7%	3.6%	9.0%	2.8%	6.9%
TIME	4.27 seconds		7.68 seconds		14.01 seconds		30.02 seconds	
HK2 BIAS	−0.0%	2.0%	0.2%	1.5%	−0.1%	2.8%	0.2%	1.2%
STD	12.1%	29.8%	10.0%	24.8%	8.8%	21.0%	7.4%	18.1%
MAD	9.6%	24.1%	8.2%	19.8%	7.0%	17.1%	5.9%	14.6%
RMSE	12.1%	29.9%	10.0%	24.8%	8.8%	21.2%	7.4%	18.1%
TIME	0.58 seconds		1.05 seconds		3.00 seconds		3.96 seconds	

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

TABLE C1B. Design 1, numerical bootstrap.

	$n = 2,500$		$n = 5,000$		$n = 10,000$		$n = 20,000$	
	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$
$c = 0.8$ COV	87.8%	89.8%	91.0%	90.7%	91.6%	93.3%	91.1%	92.5%
LEN	96.7%	99.1%	81.6%	83.2%	68.7%	68.4%	57.8%	54.7%
$c = 0.9$ COV	87.6%	89.8%	91.7%	91.5%	91.7%	94.1%	90.8%	93.5%
LEN	94.5%	97.3%	80.2%	82.5%	67.4%	68.2%	56.8%	55.1%
$c = 1.0$ COV	88.4%	90.1%	91.6%	91.1%	92.0%	93.5%	90.9%	93.2%
LEN	93.1%	96.3%	78.7%	81.8%	66.2%	68.0%	55.8%	55.5%
$c = 1.1$ COV	88.1%	89.9%	92.4%	90.9%	91.9%	93.9%	91.3%	93.2%
LEN	91.5%	94.7%	77.8%	81.1%	65.4%	67.7%	55.1%	55.7%
$c = 1.2$ COV	88.5%	89.8%	91.4%	90.6%	91.6%	93.5%	91.6%	94.1%
LEN	90.0%	93.0%	76.8%	80.0%	64.5%	67.5%	54.3%	55.7%

TABLE C1C. Design 1, $n = 1,000$.

	BIAS	STD	MAD	RMSE	COV	LEN
OY $\hat{\beta}_2$	6.4%	28.7%	22.5%	29.4%	83.5%	114.1%
$\hat{\gamma}$	−4.2%	35.0%	27.0%	35.2%	85.5%	116.3%
HK1 $\hat{\beta}_2$	−0.2%	7.6%	6.1%	7.6%	–	–
$\hat{\gamma}$	2.5%	21.7%	17.2%	21.8%	–	–
HK2 $\hat{\beta}_2$	0.2%	14.4%	11.7%	14.4%	–	–
$\hat{\gamma}$	1.0%	36.9%	29.4%	36.9%	–	–

TABLE C2A. Design 2, performance of $\hat{\beta}$ and $\hat{\gamma}$.

	$n = 2,500$		$n = 5,000$		$n = 10,000$		$n = 20,000$	
	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$
OY BIAS	2.2%	0.9%	1.9%	0.9%	0.3%	1.3%	−0.1%	1.1%
STD	19.0%	22.9%	15.0%	17.7%	11.7%	14.7%	9.2%	11.2%
MAD	15.1%	18.5%	11.9%	14.1%	9.4%	11.9%	7.3%	9.0%
RMSE	19.2%	22.9%	15.1%	17.7%	11.7%	14.8%	9.2%	11.2%
HK1 BIAS	−0.0%	2.9%	0.1%	2.6%	−0.1%	2.0%	−0.0%	1.8%
STD	5.8%	12.9%	4.2%	10.4%	3.5%	8.2%	2.7%	6.5%
MAD	4.7%	10.6%	3.4%	8.7%	2.8%	6.8%	2.2%	5.4%
RMSE	5.8%	13.3%	4.2%	10.8%	3.5%	8.5%	2.7%	6.8%
HK2 BIAS	0.2%	3.4%	0.2%	2.6%	−0.2%	2.3%	0.3%	2.0%
STD	11.9%	26.4%	10.1%	23.5%	8.9%	19.1%	6.9%	16.2%
MAD	9.6%	21.3%	8.1%	19.0%	7.2%	15.4%	5.5%	12.9%
RMSE	11.9%	26.6%	10.1%	23.6%	8.9%	19.2%	6.9%	16.3%

https://doi.org/10.1017/S0266466624000057 Published online by Cambridge University Press

TABLE C2B. Design 2, numerical bootstrap.

	$n = 2,500$		$n = 5,000$		$n = 10,000$		$n = 20,000$	
	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$	$\hat{\beta}_2$	$\hat{\gamma}$
$c = 0.8$ COV	89.4%	91.7%	90.1%	91.1%	89.2%	91.4%	91.2%	92.3%
LEN	97.2%	99.8%	81.5%	82.8%	68.8%	67.6%	57.9%	54.4%
$c = 0.9$ COV	89.8%	91.6%	90.5%	91.4%	89.7%	91.6%	91.6%	93.0%
LEN	95.2%	98.1%	80.3%	82.4%	67.5%	68.0%	57.0%	54.8%
$c = 1.0$ COV	89.6%	91.4%	91.7%	91.4%	89.0%	92.3%	92.2%	93.7%
LEN	93.4%	96.4%	78.7%	81.7%	66.5%	67.8%	56.1%	55.2%
$c = 1.1$ COV	89.1%	90.7%	91.0%	91.6%	89.0%	92.2%	91.6%	93.9%
LEN	92.0%	95.1%	77.7%	80.8%	65.4%	67.4%	55.2%	55.5%
$c = 1.2$ COV	88.8%	90.8%	91.0%	91.6%	89.4%	91.6%	91.9%	93.8%
LEN	90.5%	93.4%	76.6%	80.0%	64.6%	67.3%	54.5%	55.5%

SUPPLEMENTARY MATERIAL

Fu Ouyang and Thomas Tao Yang (2024): Supplement to "SEMIPARAMETRIC ESTIMATION OF DYNAMIC BINARY CHOICE PANEL DATA MODELS," Econometric Theory Supplementary Material. To view, please visit https://doi.org/10.1017/S0266466624000057.

REFERENCES

Abrevaya, J., & Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*, 73, 1175–1204.

Aguirregabiria, V., & Carro, J. M. (2021). Identification of average marginal effects in fixed effects dynamic discrete choice models. Preprint, arXiv:2107.06141.

Alessie, R., Hochguertel, S., & Soest, A. v. (2004). Ownership of stocks and mutual funds: A panel data analysis. *Review of Economics and Statistics*, 86, 783–796.

Al-Sadoon, M. M., Li, T., & Pesaran, M. H. (2017). Exponential class of dynamic binary choice panel data models with fixed effects. *Econometric Reviews*, 36, 898–927.

Altonji, J. G., & Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, 73, 1053–1102.

Arellano, M., & Carrasco, R. (2003). Binary choice panel data models with predetermined variables. *Journal of Econometrics*, 115, 125–157.

Arellano, M., & Honoré, B. (2001). Panel data models: Some recent developments. In J. J. Heckman, & E. Leamer (Eds.), *Handbook of econometrics*, vol. 5 (pp. 3229–3296). Elsevier.

Aristodemou, E. (2021). Semiparametric identification in panel data discrete response models. *Journal of Econometrics*, 220, 253–271.

Bartolucci, F., & Nigro, V. (2010). A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator. *Econometrica*, 78, 719–733.

Bartolucci, F., & Nigro, V. (2012). Pseudo conditional maximum likelihood estimation of the dynamic logit model for binary panel data. *Journal of Econometrics*, 170, 102–116.

Biewen, M. (2009). Measuring state dependence in individual poverty histories when there is feedback to employment status and household composition. *Journal of Applied Econometrics*, 24, 1095–1116.

Cameron, S. V., & Heckman, J. J. (1998). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political economy*, 106, 262–333.
Cameron, S. V., & Heckman, J. J. (2001). The dynamics of educational attainment for black, hispanic, and white males. *Journal of Political Economy*, 109, 455–499.
Cattaneo, M. D., Jansson, M., & Nagasawa, K. (2020). Bootstrap-based inference for cube root asymptotics. *Econometrica*, 88, 2203–2219.
Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica*, 78, 159–168.
Charlier, E. (1997). Limited dependent variable models for panel data. Technical report. Tilburg University, School of Economics and Management.
Chay, K. Y., Hoynes, H. W., & Hyslop, D. (1999). A non-experimental analysis of true state dependence in monthly welfare participation sequences. *American Statistical Association*, 9–17.
Chen, S., Khan, S, & Tang, X. (2018). Exclusion restrictions in dynamic binary choice panel data models. Technical report. Boston College, Department of Economics.
Chen, S., Khan, S., & Tang, X. (2019). Exclusion Restrictions in Dynamic Binary Choice Panel Data Models: Comment on "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors". *Econometrica*, 87, 1781–1785.
Chernozhukov, V., Fernández-Val, I., Hahn, J., & Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81, 535–580.
Chintagunta, P., Kyriazidou, E., & Perktold, J. (2001). Panel data analysis of household brand choices. *Journal of Econometrics*, 103, 111–153.
Damrongplasit, K., Hsiao, C., & Zhao, X. (2018). Health status and labour market outcome: Empirical evidence from Australia. *Pacific Economic Review*, 24, 269–292.
Dano, K. (2023). Transition probabilities and identifying moments in dynamic fixed effects logit models. Preprint, arXiv:2303.00083.
Davezies, L., D'Haultfoeuille, X., & Laage, L. (2022). Identification and estimation of average marginal effects in fixed effects logit models. Preprint arXiv:2105.00879.
Delgado, M., Rodríguez-Poo, J., & Wolf, M. (2001). Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator. *Economics Letters*, 73, 241–250.
Dobronyi, C., Gu, J., & Kim, K. I. (2021). Identification of dynamic panel logit models with fixed effects. Preprint, arXiv:2104.04590.
Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *RAND Journal of Economics*, 38, 1002–1019.
Halliday, T. J. (2008). Heterogeneity, state dependence and health. *Econometrics Journal*, 11, 499–516.
Heckman, J. J. (1981a). The incidental parameters problem and the problem of initial condition in estimating a discrete time-discrete data stochastic process. In C. F. Manski and D. L. McFadden (Eds.), *Structural analysis of discrete data and econometric applications*. MIT Press, 179–195.
Heckman, J. J. (1981b). Statistical models for discrete panel data. In C. F. Manski, & D. L. McFadden (Eds.), *Structural analysis of discrete data and econometric applications*. MIT Press, 114–178.
Hong, H., & Li, J. (2020). The numerical bootstrap. *Annals of Statistics*, 48, 397–412.
Honoré, B. E., & De Paula, Á. (2021). Identification in simple binary outcome panel data models. *Econometrics Journal*, 24, C78–C93.
Honoré, B. E., & Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68, 839–874.
Honoré, B. E., & Lewbel, A. (2002). Semiparametric binary choice panel data models without strictly exogeneous regressors. *Econometrica*, 70, 2053–2063.
Honoré, B. E., & Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 74, 611–629.
Honoré, B. E., & Weidner, M. (2020). Moment conditions for dynamic panel logit models with fixed effects. Preprint, arXiv:2005.05942.
Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60, 505–531.

Hsiao, C. (2022). *Analysis of panel data*, 4th ed. Econometric Society Monographs. Cambridge University Press.

Kerr, W. R., Lincoln, W. F., & Mishra, P. (2014). The dynamics of firm lobbying. *American Economic Journal: Economic Policy*, 6, 343–79.

Khan, S., Ponomareva, M., & Tamer, E. (2020). Identification of dynamic panel binary response models. Boston College Working Papers in Economics 979. Boston College Department of Economics.

Kim, J., & Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, 18, 191–219.

Kitazawa, Y. (2022). Transformations and moment conditions for dynamic fixed effects logit models. *Journal of Econometrics*, 229, 350–362.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65, 1335.

Lee, S. M. S., & Pun, M. C. (2006). On m out of n bootstrapping for nonstandard m-estimation with nuisance parameters. *Journal of American Statistical Association*, 101, 1185–1197.

Liu, L., Poirier, A., & Shiu, J.-L. (2023). Identification and estimation of average partial effects in semiparametric binary response panel models. Preprint, arXiv:2105.12891.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3, 205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response. *Journal of Econometrics*, 27, 313–333.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, 55, 357–362.

McFadden, D. L. (1976). Quantal choice analysis: A survey. *Annals of Economic and Social Measurement*, 5, 363–390.

Pakes, A., & Porter, J. (2016). Moment inequalities for multinomial choice with fixed effects. Technical report. National Bureau of Economic Research.

Patra, R. K., Seijo, E., & Sen, B. (2018). A consistent bootstrap procedure for the maximum score estimator. *Journal of Econometrics*, 205, 488–507.

Price, K. V., Storn, R. M., & Lampinen, J. A. (2006). *A practical approach to global optimization*. Springer-Verlag.

Seo, M. H., & Otsu, T. (2018). Local M-estimation with discontinuous criterion for dependent and limited observations. *Annals of Statistics*, 46, 344–369.

Shi, X., Shum, M., & Song, W. (2018). Estimating semiparametric panel multinomial choice models using cyclic monotonicity. *Econometrica*, 86, 737–761.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 310–313.

Storn, R. M., & Price, K. V. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11, 341–359.

Torgovitsky, A. (2019). Nonparametric inference on state dependence in unemployment. *Econometrica*, 87, 1475–1505.

van der Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes*. Springer.

Williams, B. (2019). Nonparametric identification of discrete choice models with lagged dependent variables. *Journal of Econometrics* 215, 286–304.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20, 39–54.

Yan, J., & Yoo, H. I. (2019). Semiparametric estimation of the random utility model with rank-ordered choice data. *Journal of Econometrics*, 211, 414–438.