Identification of Dynamic Panel Logit Models with Fixed Effects

Christopher Dobronyi[*] Jiaying Gu[†] Kyoo il Kim[‡]

University of Chicago *University of Toronto* *Michigan State University*

First Version: April 15, 2021, This Version: Dec 20, 2021

Abstract

We show that the identification problem for a class of dynamic panel logit models with fixed effects has a connection to the *truncated moment problem* in mathematics. We use this connection to show that the sharp identified set of the structural parameters is characterized by a set of moment equality and inequality conditions. This result provides sharp bounds in models where moment equality conditions do not exist or do not point identify the parameters. We also show that the sharp identified set of the non-parametric latent distribution of the fixed effects is characterized by a vector of its generalized moments, and that the number of moments grows linearly in T. This final result lets us point identify, or sharply bound, specific classes of functionals, without solving an optimization problem with respect to the latent distribution. We illustrate our identification result with several examples, and an empirical application on modeling children's respiratory conditions.

Keywords: Stieltjes Truncated Moment Problem, Dynamic Panel Logit Model, Fixed Effects, Moment Inequalities, Functionals of Latent Distribution

[*]Christopher Dobronyi, Kenneth C. Griffin Department of Economics, University of Chicago, 5757 S University Avenue, Chicago, Illinois 60637. Email: dobronyi@uchicago.edu.

[†]Jiaying Gu, Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario, M5S3G7, Canada. Email: jiaying.gu@utoronto.ca.

[‡]Kyoo il Kim, Department of Economics, Michigan State University, 486 W. Circle Dr, East Lansing, MI 48864, USA. Email: kyookim@msu.edu.

1 Introduction

We study the identification of a class of dynamic panel logit models with fixed effects. We show that the sharp identified set of the structural parameters is characterized by a set of moment equality and inequality conditions. Most of the literature focuses on finding moment equality conditions. Our moment inequality conditions provide sharp bounds in models without moment equality conditions, and sharpen the identified set when moment equality conditions generate multiple solutions. Our approach is analytic and easy to implement. It also provides some interesting comparisons to the sufficient statistics approach (e.g. Chamberlain (1985), Honoré and Kyriazidou (2000), Hahn (2001)) as well as the functional differencing approach (e.g. Johnson (2004), Buchinsky, Hahn, and Kim (2010), Bajari, Hahn, Hong, and Ridder (2011), Bonhomme (2012), Honoré and Weidner (2020)).

We also show that the sharp identified set of the non-parametric latent distribution of the individual fixed effects is characterized by a finite vector of its *generalized* moments. This last result allows us to *point identify* a class of functionals, including average marginal effects of the lagged choice and other counterfactual parameters (which involve the latent distribution of the fixed effects), even though the latent distribution itself is not point identified. It also allows us to characterize conditions under which functionals can be sharply bounded without solving an infinite dimensional optimization problem with respect to the latent distribution.

The dynamic panel logit model is an indispensable empirical tool for modeling repeated choices made by households, firms and individual consumers. It is commonly used, in part, because it can easily account for *permanent unobserved heterogeneity*, letting us distinguish between *true dynamics*, induced by lagged choice dependence, and *spurious dynamics*, as a result of individual heterogeneity (Heckman (1981a)). The challenges of these models are mainly due to the well known incidental parameter problem and the initial condition problem when the number of time periods T is fixed. The incidental parameter problem makes it difficult to consistently estimate structural parameters that capture the *true dynamics*. When T is fixed, it is generally not possible to treat the individual fixed effects as parameters and estimate them consistently in nonlinear panel models. Attempting to do so would also contaminate the estimation of structural parameters (Neyman and Scott (1948)). The initial condition problem states that the joint distribution of the initial value of the choices and the unobserved heterogeneity is not nonparametrically point identified, making it hard to estimate counterfactual parameters, like the average marginal effect, and other functionals of the distribution of the fixed effects (e.g., Heckman (1981b) and Wooldridge (2005a)).

There are two common approaches to deal with these challenges.[1] The (correlated) random effect approach places a restriction on the joint distribution of the initial condition and the unobserved heterogeneity through a parametric distribution, or a finite mixture model assumption (e.g., Chamberlain (1980) and Wooldridge (2005b)). If these assumptions are satisfied, then structural parameters and functionals of the latent distribution can be point identified and consistently estimated. However, if they are not satisfied, we can obtain misleading results.

The fixed effect approach is entirely non-parametric with respect to the unobserved heterogeneity. For some models, the structural parameters can be identified and consistently estimated using the *conditional maximum likelihood method*, pioneered by Andersen (1970) and Chamberlain (1985). This method involves finding a minimally sufficient statistic for the fixed effect, and constructing a partial likelihood that conditions on this statistic. By the definition of sufficiency, this partial likelihood no longer depends on the fixed effects. If this partial likelihood depends on the structural parameters, then we obtain moment equality conditions that can be used for identification and estimation. Honoré and Kyriazidou (2000) extend this approach to dynamic logit models with time varying covariates and Aguirregabiria, Gu, and Luo (2020) apply it to structural dynamic logit models in which agents make forward-looking choices. This method is easy to implement, but it does not always result in useful moment equality conditions, and even if it does, it can fail to find *all* of the relevant moment conditions.[2]

Indeed, in a recent paper by Honoré and Weidner (2020), the authors apply the functional differencing approach in Bonhomme (2012) and find new moment equality conditions for the structural parameters (in addition to those that can be found using the sufficient statistics approach) in the AR(1) dynamic panel logit model with covariates.[3] They also find moment equality conditions in models for which the sufficient statistics approach provides no moment conditions, as in the AR(2) dynamic panel logit model. The functional differencing approach is able to find more moment equalities than the sufficient statistics approach because it searches for them using the *full*

[1] For a more complete survey of the literature, we refer the readers to Arellano and Honoré (2001).

[2] There is one exception. If the likelihood of the sufficient statistics no longer depends on the structural parameters, then the conditional maximum likelihood method guarantees to utilize all relevant information on the structural parameters. In most cases, as in the dynamic panel logit model, this condition is not satisfied.

[3] For the AR(1) model, Kitazawa (2021) also derives a set of moment equality conditions using a transformation method. Neither Honoré and Weidner (2020) or Kitazawa (2021) consider moment inequalities, and they refer to moment equality conditions as moment conditions. We make the distinction here between moment equality and inequality conditions for our analysis. Bajari, Hahn, Hong, and Ridder (2011) also use a functional differencing approach to generate moment equality conditions in static discrete games with complete or incomplete information. The latent variable in their context is the equilibrium selection mechanism when multiple equilibria exist.

likelihood.

We propose a new formulation of the full likelihood. This formulation reveals a polynomial structure for the fixed effects in dynamic panel logit models. It paves the way for an algebraic approach to constructing *all* moment equality conditions for the structural parameters. This approach involves finding the basis of the left null space of a matrix that only depends on the structural parameters. When the left null space is of zero dimension, the model does not provide any moment equality conditions. Else, the resulting set of moment equality conditions is proven to coincide with the set of moment equality conditions derived using the functional differencing approach. Our results complement Honoré and Weidner (2020) by providing an algebraic way to deduce the number and the form of the moment equality conditions for the structural parameters, and address the completeness question (i.e. whether we have found *all* moment equality conditions from the model restriction) which stands as a conjecture in Honoré and Weidner (2020).

Our formulation also reveals a connection to the *truncated moment problem*, dating back to Chebyshev (1874). This connection leads to the aforementioned moment inequalities, which, with the moment equalities (if they exist) characterize the *sharp identified set* of the structural parameters. Using this result, we are able to (i) construct the sharp identified set of the structural parameters when moment equality conditions are not available, as in the AR(1) dynamic panel logit model with two periods, and (ii) rule out false roots in models where moment equality conditions solve for multiple roots, as in the AR(1) dynamic panel logit model with only a time trend covariate (see the discussion on multiple roots in Section 2.1.3 in Honoré and Weidner (2020) on the model with time trend). We further show that the model with time dummies presents a similar feature. In each of these models, the moment equality conditions pin down a finite set of parameter values that are observationally equivalent, and the moment inequality conditions can be used to rule out some of these values, and hence sharpen the identified set. We demonstrate with an empirical application the usefulness of the moment inequality information in Section 7.

It is equally important to understand the identification of the latent distribution in these models.[4] Researchers are often interested in counterfactual parameters which involve the distribution of the fixed effects. The literature on dynamic discrete choice models proposes ways to obtain the sharp identified set of the structural parameters and functionals of the latent distribution via optimization. For instance, the linear programming approach in Honoré and Tamer (2006), or the

[4]Both the conditional maximum likelihood approach in Chamberlain (1985) and the functional differencing approach by Bonhomme (2012) aim at differencing out the fixed effects to derive moment conditions for the structural parameters, and do not consider the identification of the latent distribution.

quadratic programming approach in Chernozhukov, Fernández-Val, Hahn, and Newey (2013), can be applied for this purpose. These optimization approaches search for the existence of a probability measure for the fixed effects that rationalizes the population choice probabilities for a given value of the structural parameters. This collects the set of all parameters and latent distributions that produce an observationally equivalent model. After solving for this set, we can construct sharp bounds for any integrable functional of the latent distribution of the fixed effects.

The optimization approaches, described above, are widely applicable, and can be used for models beyond those with a logistic error assumption. However, they can be challenging from a practical point-of-view, because identification is characterized using an *infinite dimensional existence problem*.[5] By focusing on dynamic panel *logit* models, we show that, the sharp identified set of the latent distribution is characterized by a finite vector of generalized moments and that the length of this vector grows only linearly in T. Using this result, for a class of functionals, we can convert the *infinite* dimensional bounding problem to a much simpler optimization problem with respect to the *finite* dimensional structural parameters.

As an example of the results described above, we show that the average marginal effect of the lagged choice in the dynamic panel logit model is a linear combination of the vector of generalized moments, which we can learn directly from the data as soon as the structural parameters are point identified. This explains why it is possible to point identify the average marginal effect even when the latent distribution itself is not point identified. This echoes the important findings in Aguirregabiria and Carro (2020), who are the first to show that the average marginal effect is point identified in a class of dynamic panel logit models. We generalize their results by providing a set of conditions on functionals, under which these functionals are point identified and provide examples of other counterfactual parameters that enjoy point identification. Moreover, while Aguirregabiria and Carro (2020) restrict their attention to models in which the structural parameters are point identified (in order to make use of a sequential identification argument), we are able to consider models in which the structural parameters are only partially identified, and sharply bound functionals using a *finite* dimensional optimization problem.

Lastly, we note that, although we focus on the dynamic logit model throughout the paper, the polynomial structure with respect to the fixed effects arises in all panel logit models including static logit models, multinomial choice logit, ordered logit model with individual fixed effects as well as

[5] Even if one can focus on the set of finite distribution for the latent variable, as pointed out by Chernozhukov, Fernández-Val, Hahn, and Newey (2013), it still requires a grid in the support of the latent variable to be computationally feasible, which may impose subtle restrictions on the set of probability measure considered.

triangular complete information logit game with market level heterogeneity (see an application of our approach in Aguirregabiria, Gu, and Mira (2021)), and hence the identification approach in this paper can naturally extend to these models.

The rest of the paper is organized as follows. Section 2 introduces the identification problem, and provides a simple example in order to illustrate our approach. Section 3 discusses our general results on the structural parameters. Section 4 includes identification results on the latent distribution and its functionals. Section 5 provides a collection of informative examples. These examples include the AR(1) and AR(2) dynamic panel logit models with and without a time trend, time dummies, and covariates. Section 6 briefly discusses estimation and inference using our identification results followed by an empirical application in Section 7. Section 8 concludes. Technical details, proofs and algebraic derivations are gathered in the Appendix.

2 Dynamic Panel Logit Model with Fixed Effects

As our baseline, we consider a dynamic panel logit model with one lag, covariates, and fixed effects, defined by:

$$Y_{it} = \mathbb{1}\{\alpha_i + \beta Y_{it-1} + X_{it}'\gamma \geq \epsilon_{it}\}, \tag{2.1}$$

where we observe outcomes $Y_{it} \in \{0, 1\}$ and covariates X_{it}, for all individuals $i = 1, \ldots, n$ over time periods $t = 0, 1, \ldots, T$. The latent variable α_i characterizes the permanent unobserved heterogeneity. This variable is allowed to have a nonparametric distribution that depends on the initial choice Y_{i0} and the covariates $X_i = (X_{i1}, \ldots, X_{iT})$. Covariates are strictly exogenous with respect to the error term ϵ_{it} and have discrete support $\mathcal{X}$ with cardinality $|\mathcal{X}|$. The error term is independently and identically distributed with respect to a standard logistic distribution. Since we focus on identification, the individual index i is dropped for the rest of the paper, unless explicitly needed. Throughout, our identification analysis is conditional on Y_0 taking a fixed value y_0.[6]

Let $\theta = \{\beta, \gamma\}$ denote the structural parameters, and let $\mathcal{Y}$ denote the set containing all possible choice histories, $\mathcal{Y} := \{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^J\}$, for which $J = 2^T$. Then, the likelihood of the choice history $\boldsymbol{y}^j$

[6] By fixing the initial choice at y_0, we mute the variation of Y_0 so that it becomes clear what is the identifying content of the model for each fixed values of y_0. When there is indeed variation in Y_0, it can provide additional identifying constraints for the structural parameters because they do not change for different values of y_0. For instance, we will take an intersection of the identified set of the structural parameters constructed using $y_0 = 0$ and $y_0 = 1$. Our results can also be applied to cases where Y_0 is not observed, but is assumed to have a degenerate distribution. This may be a reasonable assumption in certain empirical applications.

conditional on $(y_0, \boldsymbol{x}, \alpha)$ equals:

$$\mathbb{P}((Y_1, \ldots, Y_T) = \boldsymbol{y}^j \mid Y_0 = y_0, X = \boldsymbol{x}, \alpha) := \mathcal{L}_j(\alpha, \theta, \boldsymbol{x}, y_0) = \prod_{t=1}^{T} \frac{\exp(\alpha + \beta y_{t-1} + \gamma' x_t)^{y_t}}{1 + \exp(\alpha + \beta y_{t-1} + \gamma' x_t)}, \quad (2.2)$$

where $\boldsymbol{y}^j = (y_1, \ldots, y_T)$ and $\boldsymbol{x} = (x_1, \ldots, x_T)$. Integrating out the fixed effects leads to the population choice probability, denoted

$$\mathcal{P}_j = \mathbb{P}((Y_1, \ldots, Y_T) = \boldsymbol{y}^j | Y_0 = y_0, X = \boldsymbol{x}).$$

We further denote the vector $\boldsymbol{\mathcal{P}_x} = \{\mathcal{P}_1, \ldots, \mathcal{P}_J\}$. When $\gamma = 0$, this model reduces to the model in Chamberlain (1985). This model can be generalized to incorporate more than one lag.

2.1 Identification Analysis

For exposition we consider the one covariate case. The identification analysis easily extends to multiple covariates. Define $A = \exp(\alpha)$, $B = \exp(\beta)$, and $C = \exp(\gamma)$, and let $Q(A \mid y_0, \boldsymbol{x})$ denote the conditional distribution of A with support $\mathcal{A} = [0, \infty)$. Note here we have transformed α to A for convenience, because the fixed effect always enters the likelihood through the exponent function. The vector $\boldsymbol{\mathcal{P}_x}$ is assumed to be observed for our identification analysis. Let $\boldsymbol{\mathcal{L}}(A, \theta, \boldsymbol{x}, y_0)$ denote the vector that stacks $\mathcal{L}_j(A, \theta, \boldsymbol{x}, y_0)$, for $j = 1, \ldots, J$. For each tuple $(\theta, y_0, \boldsymbol{x})$, the identified set of the latent distribution of the fixed effects is the set of probability measures Q on $\mathcal{A}$ that rationalize the population choice probability $\boldsymbol{\mathcal{P}_x}$:

$$\mathcal{Q}(\theta, y_0, \boldsymbol{x}) = \Big\{ Q : \boldsymbol{\mathcal{P}_x} = \int_{\mathcal{A}} \boldsymbol{\mathcal{L}}(A, \theta, \boldsymbol{x}, y_0) dQ(A | y_0, \boldsymbol{x}) \Big\}. \quad (2.3)$$

Definition 2.1 (Identified Set)**.** *The identified set of the structural parameter θ is*

$$\Theta^* = \{\theta : \mathcal{Q}(\theta, y_0, \boldsymbol{x}) \neq \emptyset, \text{ for all } \boldsymbol{x} \in \mathcal{X}\}.$$

Moreover, the joint identified set of the structural parameter θ and the latent distribution is

$$\mathcal{I}^*(y_0, \boldsymbol{x}) = \{(\theta, Q) : \theta \in \Theta^* \text{ and } Q \in \mathcal{Q}(\theta, y_0, \boldsymbol{x})\}.$$

If Θ^* is a singleton, then θ is point identified, and the true distribution of the fixed effects, denoted $Q_0(A \mid y_0, \boldsymbol{x})$, is known to be a member of $\mathcal{Q}(\theta_0, y_0, \boldsymbol{x})$ where θ_0 denotes the true value of θ. The question of whether a point θ belongs to the identified set Θ^* can be viewed as an *infinite* *dimensional existence problem*—it asks whether there exists a probability measure Q such that the observed vector of choice probabilities can be rationalized by the model given θ. We now show that we can reduce this *infinite dimensional existence problem* to a *finite* set of moment equality and inequality conditions, for each $\boldsymbol{x} \in \mathcal{X}$. Furthermore, for each $\theta \in \Theta^*$, the set $\mathcal{Q}(\theta, y_0, \boldsymbol{x})$ defined in (2.3), can be equivalently characterized by a finite vector of generalized moments of A.

2.2 Simple Case: Two Time Periods without Covariates

We first present this result for the simple case in which $T = 2$ and $\gamma = 0$ to establish intuition. This simple case reveals a fundamental connection between the identification problem and the *truncated moment problem* in mathematics. We use this simple case to motivate our general identification results in Sections 3 and 4. This simple case is also interesting by itself. Honoré and Weidner (2020) have shown that there are no moment equality conditions for the structural parameter. We show that the model still provides information about the structural parameters through a finite set of moment inequalities that define the sharp identified set. To the best of our knowledge, our paper is the first to establish this sharp identification result for this model.

For exposition, we fix $y_0 = 0$.[7] In this setting, the likelihood vector can be denoted $\boldsymbol{\mathcal{L}}(A, \beta)$ and written as:[8]

$$\boldsymbol{\mathcal{L}}(A, \beta) = G(\beta) \begin{pmatrix} 1 \\ A \\ A^2 \\ A^3 \end{pmatrix} \frac{1}{g(A, \beta)}, \tag{2.4}$$

where $g(A, \beta) = (1 + A)^2(1 + AB)$ and the matrix $G(\beta)$ is defined by:

$$G(\beta) = \begin{pmatrix} 1 & B & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & B & 0 \\ 0 & 0 & B & B \end{pmatrix}.$$

Therefore, in this model, we can write:

$$\boldsymbol{\mathcal{P}} = (p_0, p_1, p_2, p_3)' = G(\beta) \int_{\mathcal{A}} \begin{pmatrix} 1 & A & A^2 & A^3 \end{pmatrix}' d\bar{Q}(A \mid \beta), \tag{2.5}$$

where $\boldsymbol{\mathcal{P}}$ is the vector of population choice probabilities and $d\bar{Q}(A \mid \beta) = \frac{1}{g(A,\beta)} dQ(A)$. Since

[7] The same results derived can be extended to the case where $y_0 = 1$.

[8] The elements in the set $\mathcal{Y}$ is ordered as: $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$. See details in Appendix A.4.

$1/g(A, \beta)$ is bounded for all $A \in \mathcal{A}$, the measure $\bar{Q}(A \mid \beta)$ is a finite positive Borel measure on $\mathcal{A}$. It is easy to check that $G(\beta)$ is of full rank unless $\beta = 0$ (which we rule out).[9]

There are several features of the formulation in (2.4) that are worth mentioning. The choice of $g(A, \beta)$ is natural. Since $\mathcal{L}_j(A, \beta)$ is a ratio of polynomials of A, we can choose $g(A, \beta)$ to be a polynomial of A of the smallest degree for which the product $\mathcal{L}_j(A, \beta)g(A, \beta)$ is a polynomial of A, for all $j = 1, \ldots, J$. Doing so leads to a matrix $G(\beta)$ that does not depend on the fixed effects.

Since $G(\beta)$ is of full rank, we obtain:

$$\boldsymbol{r}(\beta) = (r_0(\beta), \ldots, r_3(\beta))' := G(\beta)^{-1}\boldsymbol{\mathcal{P}} = \int_{\mathcal{A}} \begin{pmatrix} 1 & A & A^2 & A^3 \end{pmatrix}' d\bar{Q}(A \mid \beta).$$

Notice that, for a given β, the vector $\boldsymbol{r}(\beta)$ is observed (since $\boldsymbol{\mathcal{P}}$ denotes the observed vector of choice probabilities and $G(\beta)^{-1}$ is a known matrix). The right-hand side of this equation is a vector of moments up to order 3 with respect to the measure $\bar{Q}(A \mid \beta)$. Hence, the question of whether a particular value of β belongs to the identified set translates into the question of whether there exists a finite positive Borel measure for which the vector $\boldsymbol{r}(\beta)$ can be written as a vector of moments up to order 3. This result reveals the fundamental connection to the *truncated moment problem*, which we alluded to in the previous section. In particular, one of the questions studied in the literature on the *truncated moment problem* is whether there exists a Borel measure that rationalizes a finite sequence of numbers as its moments.

To this end, we define the moment space of any positive Borel measure μ on $[0, \infty)$ to be:

$$\mathcal{M}_K = \left\{ \boldsymbol{c} \in \mathbb{R}^{K+1} : \text{ there exists } \mu \text{ such that } c_k = \int_0^{+\infty} A^k d\mu(A), \text{ for all } k = 0, 1, \ldots, K \right\}.$$

With this definition, we can write $\Theta^* = \{\beta : \boldsymbol{r}(\beta) \in \mathcal{M}_3\}$ in this simple case. The unique geometric structure of the moment space $\mathcal{M}_K$ leads to the following theorem.

Theorem 2.1. *For the dynamic logit model in (2.1) with $T = 2$ and $\gamma = 0$, the value $\beta \in \Theta^*$ if and only if $\sum_{j=0}^{3} \eta_j r_j(\beta) \geq 0$, for every non-trivial real-valued sequence of coefficients $\{\eta_j\}_{j=0}^{3}$ such that $\sum_{j=0}^{3} \eta_j A^j \geq 0$, for all $A \in [0, \infty)$.*

Theorem 2.1 is Theorem 9.1 in Karlin and Studden (1966), applied to our context. The key insight is that the dual cone[10] of the moment space, which is itself a convex cone, can be identified

[9] If $\beta = 0$, then we have a static logit model, also known as the Rasch model by Rasch (1961). Identification of this model is well understood. See for instance Cressie and Holland (1983). For the static model with T periods, we also have a polynomial formulation of the full likelihood, although it will involve polynomial of A up to order T. We can easily test whether $\beta = 0$. For instance for $T = 2$ without covariates, $\beta = 0$ if and only if the choice probability of $(1, 0)$ equals the choice probability of $(0, 1)$.

[10] For a convex cone $\mathcal{C}$ contained in $\mathbb{R}^d$, the dual cone $\mathcal{C}^+ = \{\boldsymbol{\lambda} \in \mathbb{R}^d \mid \boldsymbol{\lambda}'\boldsymbol{c} \geq 0 \text{ for all } \boldsymbol{c} \in \mathcal{C}\}$.

as the space of non-negative polynomials of A up to degree K. Now we discuss the implication of Theorem 2.1. Every non-negative polynomial of A with an odd degree has a represention with the form:[11]

$$\sum_{j=0}^{2m+1} \eta_j A^j = Af^2(A) + q^2(A),$$

for all $A \in [0, \infty)$, where $f(A)$ and $q(A)$ are polynomials. In our context, we have $m = 1$, and $f(A)$ and $q(A)$ are polynomials of at most degree 1. Therefore, we can write $f(A) = \xi_0 + \xi_1 A$ and $q(A) = \lambda_0 + \lambda_1 A$, for any coefficients (ξ_0, ξ_1) and (λ_0, λ_1) in $\mathbb{R}^2$ such that:

$$\sum_{j=0}^{3} \eta_j A^j = A(\xi_0 + \xi_1 A)^2 + (\lambda_0 + \lambda_1 A)^2 \geq 0.$$

Retrieving the coefficient η_j and the condition $\sum_{j=0}^{3} \eta_j r_j(\beta) \geq 0$ leads to:

$$\lambda_0^2 r_0(\beta) + 2\lambda_0\lambda_1 r_1(\beta) + \lambda_1^2 r_2(\beta) + \xi_0^2 r_1(\beta) + 2\xi_0\xi_1 r_2(\beta) + \xi_1^2 r_3(\beta) \geq 0,$$

which can be equivalently stated as:

$$\begin{pmatrix} \lambda_0 & \lambda_1 \end{pmatrix} \begin{pmatrix} r_0(\beta) & r_1(\beta) \\ r_1(\beta) & r_2(\beta) \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \end{pmatrix} + \begin{pmatrix} \xi_0 & \xi_1 \end{pmatrix} \begin{pmatrix} r_1(\beta) & r_2(\beta) \\ r_2(\beta) & r_3(\beta) \end{pmatrix} \begin{pmatrix} \xi_0 \\ \xi_1 \end{pmatrix} \geq 0,$$

for *any* coefficients (λ_0, λ_1) and (ξ_0, ξ_1) in $\mathbb{R}^2$.

This last condition, as we show later in Theorems 2.2 and 2.3 (in a more general form), boils down to checking the non-negativity of the two square matrices above, defined using the elements of $\boldsymbol{r}(\beta)$. These are known as Hankel matrices in the truncated moment problem literature. Since the non-negativity of a square matrix is equivalent to all of its principal minors being non-negative, and $\boldsymbol{r}(\beta)$ is a linear combination of choice probabilities with coefficients depend on β (recall $\boldsymbol{r}(\beta) = G(\beta)^{-1}\boldsymbol{\mathcal{P}}$), we obtain *moment inequalities* for β. In fact, these moment inequalities characterize *all* of the information about β implied by this model.

[11] Every non-negative polynomial up to degree K, $\sum_{j=0}^{K} \eta_j A^j$, over $[0, \infty)$ can be written in the form: (i) if $K = 2m+1$ (odd case), then $\sum_{j=0}^{2m+1} \eta_j A^j = Af^2(A) + q^2(A)$ where $f(A)$ and $q(A)$ are polynomial functions of A up to order m. If $K = 2m$ (even case), then $\sum_{j=0}^{2m} \eta_j A^j = f^2(A) + Aq^2(A)$ where $f(A)$ are polynomials of A of at most order m and $q(A)$ is a polynomial of A up to order $m-1$. See Karlin and Studden (1966). We are in the odd case here because we have moments up to order 3.

Given the form of $G(\beta)$ in this example, we have:

$$\boldsymbol{r}(\beta) := \begin{pmatrix} r_0(\beta) \\ r_1(\beta) \\ r_2(\beta) \\ r_3(\beta) \end{pmatrix} = \frac{1}{B-1} \begin{pmatrix} (B-1)p_0 - B^2 p_1 + B p_2 \\ B p_1 - p_2 \\ -p_1 + p_2 \\ p_1 - p_2 + \frac{B-1}{B} p_3 \end{pmatrix}$$

where, once again, B denotes $\exp(\beta)$. Therefore, the non-negativity of the two Hankel matrices is equivalent to the following set of inequalities:

$$\begin{aligned} & r_j(\beta) \geq 0, \quad j = 0, \ldots, 3. \\ & r_0(\beta) r_2(\beta) - r_1(\beta)^2 \geq 0 \\ & r_1(\beta) r_3(\beta) - r_2(\beta)^2 \geq 0 \end{aligned}$$

We derive the analytical sharp bound on β from these inequalities in Appendix A.4 and demonstrate that the bound on β can be quite informative even though we only observe two periods of choices.

Furthermore, given the reformulation in (3.1), for each $\beta \in \Theta^*$, the set $\mathcal{Q}(\beta, y_0)$ defined in (2.3) can now be written:

$$\mathcal{Q}(\beta, y_0) = \left\{ Q : \boldsymbol{r}(\beta) = \int_{\mathcal{A}} \begin{pmatrix} 1 & A & A^2 & A^3 \end{pmatrix}' \frac{1}{g(A, \beta)} dQ \right\}.$$

We refer to the vector, $\int_{\mathcal{A}} (1/g, A/g, A^2/g, A^3/g) dQ$, as the generalized moment vector of the random variable A. Any distribution Q whose generalized moments coincide with $\boldsymbol{r}(\beta)$ is an element of $\mathcal{Q}(\beta, y_0)$. Since functions in the vector $\{1/g, A/g, A^2/g, A^3/g\}$ are linearly independent with support $\mathcal{A}$, which we prove more generally in Lemma 3.1, the generalized moment vector is all that we can learn about the latent distribution of the unobserved heterogeneity, for each value of β in Θ^*.

The polynomial structure in this simple example preserves in the dynamic panel logit model, both with and without covariates, for any finite T. Before we discuss the general result in the next section, we first present general results on the truncated moment problem.

2.3 Results on the Truncated Moment Problem

The *moment problem* poses the question: Is a sequence of real numbers equal to the sequence of moments associated with some finite Borel measure supported on some set $\mathbb{K}$? When the sequence of real numbers is an infinite sequence, we have the *full moment problem*. Else, we have the *truncated moment problem*. When $\mathbb{K} = [0, \infty)$, as in our context, the truncated moment problem becomes the *truncated Stieltjes moment problem*. We refer the readers to the authoritative treatments of the moment problem in Karlin and Studden (1966) and Krein and Nudelman (1977).

In our context, the finite sequence of real numbers is the vector $\boldsymbol{r}(\beta)$, and we have the following general theorem. For any $m \times n$ matrix $\boldsymbol{A}$, define $\text{Range}(\boldsymbol{A}) = \{\boldsymbol{Au}, \boldsymbol{u} \in \mathbb{R}^n\}$. The symbol $\succeq 0$ represents a square matrix being positive semidefinite.

Theorem 2.2 (Truncated Stieltjes Moment Problem (Odd Case))**.** *Let* $\boldsymbol{r} = \{r_0, r_1, \ldots, r_m\} \in \mathbb{R}^{m+1}$ *denote a finite dimensional vector. If* m *is odd (i.e.* $m = 2k+1$*), define the following matrices:*

$$\boldsymbol{H_k}(\boldsymbol{r}) = \begin{pmatrix} r_0 & r_1 & \cdots & r_k \\ r_1 & r_2 & \cdots & r_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ r_k & r_{k+1} & \cdots & r_{2k} \end{pmatrix} \quad \textit{and} \quad \boldsymbol{B_k}(\boldsymbol{r}) = \begin{pmatrix} r_1 & r_2 & \cdots & r_{k+1} \\ r_2 & r_3 & \cdots & r_{k+2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k+1} & r_{k+2} & \cdots & r_{2k+1} \end{pmatrix}$$

Then $\boldsymbol{r} \in \mathcal{M}_{2k+1}$ *if and only if* $\boldsymbol{H_k}(\boldsymbol{r}) \succeq 0$*,* $\boldsymbol{B_k}(\boldsymbol{r}) \succeq 0$ *and* $\{r_{k+1}, r_{k+2}, \ldots, r_{2k+1}\}$ *is in Range*$(\boldsymbol{H_k}(\boldsymbol{r}))$*.*

Theorem 2.3 (Truncated Stieltjes Moment Problem (Even Case))**.** *Let* $\boldsymbol{r} = \{r_0, r_1, \ldots, r_m\} \in \mathbb{R}^{m+1}$ *denote a finite dimensional vector. If* m *is even (i.e.* $m = 2k$*), define the following matrices:*

$$\boldsymbol{H_k}(\boldsymbol{r}) = \begin{pmatrix} r_0 & r_1 & \cdots & r_k \\ r_1 & r_2 & \cdots & r_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ r_k & r_{k+1} & \cdots & r_{2k} \end{pmatrix} \quad \textit{and} \quad \boldsymbol{B_{k-1}}(\boldsymbol{r}) = \begin{pmatrix} r_1 & r_2 & \cdots & r_k \\ r_2 & r_3 & \cdots & r_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ r_k & r_{k+1} & \cdots & r_{2k-1} \end{pmatrix}$$

Then $\boldsymbol{r} \in \mathcal{M}_{2k}$ *if and only if* $\boldsymbol{H_k}(\boldsymbol{r}) \succeq 0$*,* $\boldsymbol{B_{k-1}}(\boldsymbol{r}) \succeq 0$ *and* $\{r_{k+1}, r_{k+2}, \ldots, r_{2k}\}$ *is in Range*$(\boldsymbol{B_{k-1}}(\boldsymbol{r}))$*.*

Proof. See Curto and Fialkow (1991). ■

In Theorems 2.2 and 2.3, $\boldsymbol{H_k}(\boldsymbol{r})$ is a *Hankel matrix*, and $\boldsymbol{B_k}(\boldsymbol{r})$ and $\boldsymbol{B_{k-1}}(\boldsymbol{r})$ are localized verions of $\boldsymbol{H_k}(\boldsymbol{r})$ (in which the moments are shifted by one index). The condition on the range is only relevant when $\boldsymbol{H_k}(\boldsymbol{r})$ is singular. Hankel matrices have a recursive structure, as shown by Curto and Fialkow (1991). Therefore, if a Hankel matrix is singular and positive, then its entries are uniquely recursively determined by elements in a non-singular sub-Hankel matrix (of a smaller dimension). This translates into the range condition of the relevant matrix. See an explicit discussion of the range condition for model (2.1) with $T = 2$ and $\gamma = 0$ in Section A.4.

3 General Results for the AR(1) Model

With the results from the dynamic panel logit model in which $T = 2$ and $\gamma = 0$ in hands, we now generalize. For model (2.1) with $T \geq 2$ and $\theta = (\beta, \gamma)$, there exists a polynomial function $g(A, \theta, \boldsymbol{x}, y_0)$ of order $2T - 1$ such that:

$$\boldsymbol{\mathcal{L}}(A, \theta, \boldsymbol{x}, y_0) = G(\theta, \boldsymbol{x}) \begin{pmatrix} 1 \\ A \\ \vdots \\ A^{2T-1} \end{pmatrix} \frac{1}{g(A, \theta, \boldsymbol{x}, y_0)}, \tag{3.1}$$

where $G(\theta, \boldsymbol{x})$ is a matrix of dimension $2^T \times 2T$. See Appendix A.2 for an explicit formulation of $g(A, \theta, \boldsymbol{x}, y_0)$. Therefore, the population choice probabilities have the following representation:

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = \int_{\mathcal{A}} \boldsymbol{\mathcal{L}}(A, \theta, \boldsymbol{x}, y_0) dQ(A|y_0, \boldsymbol{x}) = G(\theta, \boldsymbol{x}) \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^{2T-1} \end{pmatrix}' d\bar{Q}(A|y_0, \theta, \boldsymbol{x}) \tag{3.2}$$

where $d\bar{Q}(A|y_0, \theta, \boldsymbol{x}) = \frac{1}{g(A,\theta,\boldsymbol{x},y_0)} dQ(A|y_0, \boldsymbol{x})$. Since $1/g(A, \theta, \boldsymbol{x}, y_0)$ is bounded everywhere on $\mathcal{A}$, the measure $\bar{Q}(A|y_0, \theta, \boldsymbol{x})$ is a positive Borel measure supported on $\mathcal{A}$. We further define the set of vectors:

$$\boldsymbol{M}_{\boldsymbol{x}} = \{\boldsymbol{v}_{\boldsymbol{x}} \in \mathbb{R}^{2^T} : \boldsymbol{v}_{\boldsymbol{x}}' G(\theta, \boldsymbol{x}) = 0\}, \tag{3.3}$$

where the vector $\boldsymbol{v}_{\boldsymbol{x}}$ is implicitly a function of the parameter θ and $\boldsymbol{x}$. By construction, $\boldsymbol{v}_{\boldsymbol{x}}$ is a member of the left null space of $G(\theta, \boldsymbol{x})$. Both $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ and $G(\theta, \boldsymbol{x})$ (and hence $\boldsymbol{v}_{\boldsymbol{x}}$) implicitly depend on the fixed value of y_0.

3.1 Identification of Structural Parameters

We now state our main result:

Theorem 3.1. *If $G(\theta, \boldsymbol{x})$ is full rank, then $\theta \in \Theta^*$ if and only if the following conditions hold:*

(a) For all $\boldsymbol{x} \in \mathcal{X}$, we have $\boldsymbol{v}_{\boldsymbol{x}}' \boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = 0$ for all $\boldsymbol{v}_{\boldsymbol{x}} \in \boldsymbol{M}_{\boldsymbol{x}}$.

(b) For all $\boldsymbol{x} \in \mathcal{X}$, we have $\boldsymbol{r}(\theta, \boldsymbol{x}) \in \mathcal{M}_{2T-1}$, where $\boldsymbol{r}(\theta, \boldsymbol{x}) = H(\theta, \boldsymbol{x}) \boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ and $H(\theta, \boldsymbol{x})$ is a matrix of dimension $2T \times 2^T$ such that $H(\theta, \boldsymbol{x}) G(\theta, \boldsymbol{x}) = I_{2T}$.

The requirement on $G(\theta, \boldsymbol{x})$ being full rank can be checked in each specific model. Theorem 3.1 shows that the identification of the structural parameters θ can be formulated as a set of *moment equalities* (from condition (a)) and *moment inequalities* (from condition (b)).

The moment equality conditions take the form $\mathbb{E}[\boldsymbol{v}_{\boldsymbol{x}}'\mathbb{Y}|Y_0 = y_0, X = \boldsymbol{x}] = \boldsymbol{v}_{\boldsymbol{x}}'\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = 0$, for each $\boldsymbol{x} \in \mathcal{X}$, where $\mathbb{Y}$ is a vector of length 2^T with its elements being $\mathbb{1}\{(Y_1, \ldots, Y_T) = \boldsymbol{y}^j\}$ for $\boldsymbol{y}^j \in \mathcal{Y}$. This is due to the fact that $\boldsymbol{v}_{\boldsymbol{x}}'G(\theta, \boldsymbol{x}) = 0$ implies $\boldsymbol{v}_{\boldsymbol{x}}'\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = 0$ given our representation in (3.2). The form of these moment equalities can be found as the basis of the left null space of $G(\theta, \boldsymbol{x})$, defined in $\boldsymbol{M}_{\boldsymbol{x}}$.[12]

The moment inequalities arise from the condition that $\boldsymbol{r}(\theta, \boldsymbol{x})$ has to be a moment sequence, and the fact that $\boldsymbol{r}(\theta, \boldsymbol{x})$ is a linear combination of the elements in $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$. Through the necessary and sufficient conditions of the *truncated moment problem* in Theorem 2.2 or Theorem 2.3, we impose non-negativity on the two Hankel matrices formed from elements in $\boldsymbol{r}(\theta, \boldsymbol{x})$, which is equivalent to all of its principal minors being non-negative. Since the vector $\boldsymbol{r}(\theta, \boldsymbol{x})$ is of length $2T$, the two Hankel matrices are both of dimension $T \times T$, which then leads to a finite set of principle minors. These constraints on $\boldsymbol{r}(\theta, \boldsymbol{x})$ lead to the inequality conditions for θ.

For the construction of $\boldsymbol{r}(\theta, \boldsymbol{x})$, in principle, the matrix $H(\theta, \boldsymbol{x})$ can always be chosen to be $(G(\theta, \boldsymbol{x})'G(\theta, \boldsymbol{x}))^{-1}G(\theta, \boldsymbol{x})'$. In Appendix A.3, we discuss an alternative for ease of implementation. The inequalities boil down to checking the non-negativity of two square matrices involving elements in $\boldsymbol{r}(\theta, \boldsymbol{x})$. While all of the examples in this paper have a matrix $G(\theta, \boldsymbol{x})$ with full column rank, our results can generalize to models for which $G(\theta, \boldsymbol{x})$ is not of full rank. See the discussion in Appendix A.3.

3.2 Example: Three Periods with a Covariate

We now give the example of $T = 3$ with a scalar covariate as an application of Theorem 3.1. This example allows us to compare our results with the results in Honoré and Kyriazidou (2000) as well as those in Honoré and Weidner (2020). Using the reformulation in (3.2), we have the following representation:

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = \int_{\mathcal{A}} G(\theta, \boldsymbol{x}) \begin{pmatrix} 1 & A & \cdots & A^5 \end{pmatrix}' \frac{1}{g(A, \theta, \boldsymbol{x}, y_0)} dQ(A \mid \theta, y_0, \boldsymbol{x}),$$

with $g(A, \theta, \boldsymbol{x}, y_0) = (1 + AC^{x_1})(1 + AC^{x_2})(1 + AC^{x_3})(1 + ABC^{x_2})(1 + ABC^{x_3})$ when $y_0 = 0$. The vector $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ has elements $\mathbb{P}((Y_1, \ldots, Y_T) = \boldsymbol{y}|Y_0 = y_0, X = \boldsymbol{x})$ with $\boldsymbol{y} \in \mathcal{Y}$. The elements in the set $\mathcal{Y}$ are ordered as: $\{(0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}$. We label

[12] With the symbolic tools in Matlab, the set $\mathcal{M}_{\boldsymbol{x}}$ can be easily constructed. An earlier working paper version of Buchinsky, Hahn, and Kim (2010) shows that using Johnson (2004) characterization of semiparametric information bound for discrete choice models one can derive the moment equality conditions for a few examples of dynamic panel logit models with fixed effects with $T \leq 3$. They, however, do not consider a more general dynamic panel logit specification like model (2.1) with covariates, or the moment inequalities that we develop in this paper.

$\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = (p_0, \ldots, p_7)'$. For the complete form of $G(\theta, \boldsymbol{x})$ see Appendix A.6.

The left null space of the matrix $G(\theta, \boldsymbol{x})$ is spanned by the following two vectors:

$$v_1 = \begin{pmatrix} 0 \\ C^{x_3}(B-1) \\ C^{x_1}(1 - BC^{x_3-x_2}) \\ C^{x_1}(1 - C^{x_2-x_3}) \\ BC^{x_3}(1 - C^{x_3-x_2}) \\ B(C^{x_3} - C^{x_2}) \\ 0 \\ 0 \end{pmatrix} \text{ and } v_2 = \begin{pmatrix} 0 \\ C^{x_3-x_1}(BC^{x_2} - C^{x_3}) \\ C^{x_3}(1-B) \\ C^{x_3} - C^{x_2} \\ -BC^{x_3-x_1}(C^{x_3} - C^{x_2}) \\ 0 \\ C^{x_3} - C^{x_2} \\ 0 \end{pmatrix}. \tag{3.4}$$

We then have $v_k' \boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = 0$ $(k = 1, 2)$, or equivalently $\mathbb{E}[v_k' \mathbb{Y} | Y_0 = 0, X = \boldsymbol{x}] = 0$ $(k = 1, 2)$ as our moment equality restrictions. We further note that using linear combinations of v_1 and v_2 leads to the moment conditions in Honoré and Weidner (2020) for the same model. In particular, for $y_0 = 0$, we have:

$$\begin{aligned} m_1 &\equiv \frac{1}{B(C^{x_2} - C^{x_3})} v_1 - \frac{C^{x_1}}{BC^{x_3}(C^{x_2} - C^{x_3})} v_2 = \Big(0, \quad -1, \quad C^{x_1-x_2}, \quad 0, \quad C^{x_3-x_2} - 1, \quad -1, \quad C^{x_1-x_3}/B, \quad 0\Big)', \\ m_2 &\equiv -\frac{C^{x_2}}{C^{x_1}(C^{x_2} - C^{x_3})} v_1 + \frac{1}{C^{x_2} - C^{x_3}} v_2 = \Big(0, \quad C^{x_3-x_1}, \quad -1, \quad C^{x_2-x_3} - 1, \quad 0, \quad BC^{x_2-x_1}, \quad -1, \quad 0\Big)'. \end{aligned} \tag{3.5}$$

and note $m_k' \boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = 0$ $(k = 1, 2)$ are the two moment conditions in Honoré and Weidner (2020) for the same model. A similar derivation can be done for $y_0 = 1$.

On the other hand, if we were to impose the restriction $x_2 = x_3$ as assumed in Honoré and Kyriazidou (2000), the basis for the left null space of $G(\theta, \boldsymbol{x})$ leads to the following two vectors:

$$\begin{aligned} v_1 &= \begin{pmatrix} 0 & -C^{x_2-x_1} & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}', \\ v_2 &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -BC^{x_2-x_1} & 1 & 0 \end{pmatrix}'. \end{aligned}$$

Whenever $x_1 \neq x_2$, the implied moment equalities have a unique solution:

$$\beta = \log(p_1) - \log(p_2) - \log(p_5) + \log(p_6) \text{ and } \gamma = \frac{\log(p_1) - \log(p_2)}{x_1 - x_2},$$

where $p_j = p_j(\boldsymbol{x}) = p_j(x_1, x_2, x_2)$. These moment equalities coincide with the moment conditions in Honoré and Kyriazidou (2000).

In this example, the two moment equality conditions, based on (3.4) or equivalently (3.5), alone point identifiy the parameters θ. If the model is correctly specified, then the moment inequalities must hold for all values of θ that satisfy the two moment conditions; if they do not, then the logit model is mis-specified.

Moment equality conditions do not always secure point identification. We discuss two such cases in Section 5 for model (2.1): one with a time trend variable, and one with time dummies. In these models, we illustrate the use of moment inequality conditions to sharpen the identified set.

3.2.1 Relationship with Results in Honoré and Weidner (2020)

Honoré and Weidner (2020) derive the moment equality conditions for θ using the functional differencing approach in Bonhomme (2012). To be precise, let us define the set:

$$\kappa_{\boldsymbol{x}} = \{\boldsymbol{m}_{\boldsymbol{x}} \in \mathbb{R}^{2^T} : \boldsymbol{m}_{\boldsymbol{x}}' \boldsymbol{\mathcal{L}}(A, \theta, \boldsymbol{x}, y_0) = 0, \forall A \in \mathcal{A}\}. \tag{3.6}$$

Their moment conditions consist of $\mathbb{E}[\boldsymbol{m}_{\boldsymbol{x}}' \mathbb{Y} | Y_0 = y_0, X = \boldsymbol{x}] = \boldsymbol{m}_{\boldsymbol{x}}' \boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = 0$, for all $\boldsymbol{m}_{\boldsymbol{x}} \in \kappa_{\boldsymbol{x}}$ and all $\boldsymbol{x} \in \mathcal{X}$. We now prove that the set $\boldsymbol{M}_{\boldsymbol{x}}$, defined in (3.3), coincides with the set $\kappa_{\boldsymbol{x}}$, for all $\boldsymbol{x} \in \mathcal{X}$.

Lemma 3.1. *The functions in* $\mathcal{V}_{\theta, \boldsymbol{x}, y_0}(A) = \left\{ \frac{1}{g(A, \theta, \boldsymbol{x}, y_0)}, \frac{A}{g(A, \theta, \boldsymbol{x}, y_0)}, \cdots, \frac{A^{2T-1}}{g(A, \theta, \boldsymbol{x}, y_0)} \right\}$ *are linearly independent.*

Given the formulation in (3.1) for $\boldsymbol{\mathcal{L}}$, the vectors in the set $\kappa_{\boldsymbol{x}}$ are those such that $\boldsymbol{m}_{\boldsymbol{x}}' G(\theta, \boldsymbol{x})$ are orthogonal to the vector $\mathcal{V}_{\theta, \boldsymbol{x}, y_0}(A)$ for all $A \in \mathcal{A}$. Lemma 3.1 states that the set of functions in $\mathcal{V}_{\theta, \boldsymbol{x}, y_0}$ spans $\mathbb{R}^{2T}$, which implies that $\boldsymbol{m}_{\boldsymbol{x}}' G(\theta, \boldsymbol{x}) = 0$ for all $\boldsymbol{m}_{\boldsymbol{x}} \in \kappa_{\boldsymbol{x}}$. This observation leads to the next Theorem.

Theorem 3.2. $\boldsymbol{M}_{\boldsymbol{x}} = \kappa_{\boldsymbol{x}}$ *for all* $\boldsymbol{x} \in \mathcal{X}$.

Theorem 3.2 proves that the moment equality conditions found by our approach coincide with the moment conditions in Honoré and Weidner (2020). Our moment equality conditions can always be written as linear combinations of the moment conditions in Honoré and Weidner (2020), and vice versa.

Given the equivalence in Theorem 3.2, our results provide an algebraic foundation for the results in Honoré and Weidner (2020). In addition, they use numerical evidence to conjecture there are $2^T - 2T$ moment conditions, for each given $\boldsymbol{x}$. This result can be easily verified in our framework: Provided that $G(\theta, \boldsymbol{x})$ is of full rank, the left null space of $G(\theta, \boldsymbol{x})$ is of dimension $2^T - 2T$, for any $T \geq 2$. As a result, we expect to have $2^T - 2T$ linearly independent vectors that form a basis for this space, where each vector serves as a moment equality condition for identifying θ. This feature can also be used to explain why there are no moment equality conditions available when $T = 2$, an impossibility result established in Honoré and Weidner (2020). Naturally, when $T = 2$, $G(\theta, \boldsymbol{x})$ is a 4×4 full rank square matrix, implying that its left null space is of zero dimension, and that there does not exist a vector $\boldsymbol{v_x}$ for which $\boldsymbol{v}'_{\boldsymbol{x}} G(\theta, \boldsymbol{x}) = 0$ (with the exception of the null vector). In this case, all of the identifying content of θ is contained in the moment inequality conditions that we characterize in Theorem 3.1.

3.3 Connection with the CMLE Approach

The conditional maximum likelihood approach uses sufficient statistics to factorize the likelihood into a component that depends on the fixed effects, and a component that does not. For instance, in model (2.1) without covariates, we can write:

$$\mathcal{P}(\boldsymbol{y} \mid y_0, \theta) = \mathcal{P}(\boldsymbol{y} \mid S(\boldsymbol{y}), \theta) \int_{\mathcal{A}} \mathcal{P}(S(\boldsymbol{y}) \mid A, \theta) dQ(A \mid y_0), \tag{3.7}$$

for each $\boldsymbol{y} = \{y_1, \ldots, y_T\} \in \{0, 1\}^T$, where $S(\boldsymbol{y})$ is a sufficient statistic for A. In this model, we can use $S(\boldsymbol{y}) = \{y_0, \sum_{t=1}^{T-1} y_t, y_T\}$. After factorizing the likelihood, we can derive moment equality conditions using the first component in (3.7). This procedure leads to point identification in this particular model as soon as $T \geq 3$. The *conditional maximum likelihood estimator* (CMLE) solves the empirical counterpart of the system of equations implied by these moment conditions.

While this procedure leads to a $\sqrt{n}$-consistent estimator, the second component in (3.7) depends on θ, leaving us with an interesting question: Is there useful information in the second component that is being thrown away by the CMLE approach? This question is related to an open puzzle. Specifically, Hahn (2001) shows that the CMLE does not achieve the semiparametric efficiency bound (when $T = 3$), suggesting that there might exist an estimator that is asymptotically more efficient, but no such estimator has been found to date.[13] We use our framework to revisit this puzzle. In particular, we apply our methodology to the second component in (3.7) in order to determine whether it contains any useful information about θ.

Consider model (2.1) without covariates given $T \geq 3$ and $y_0 = 0$.[14] In this model, the support of $S(\boldsymbol{y})$ contains $2T$ points since $\sum_{t=1}^{T-1} y_t \in \{0, 1, \ldots, T-1\}$ and $y_T \in \{0, 1\}$. Label these points

[13] Hahn (2001) also makes use of the polynomial structure of the dynamic panel logit model in his analysis of the semiparametric efficiency for the CMLE.

[14] We fix $y_0 = 0$ to be consistent with Hahn (2001).

$\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{2T}$. The likelihood associated with the sufficient statistic $S(\boldsymbol{y})$ is:

$$\boldsymbol{\mathcal{L}}^S(A, \theta, y_0) = \Big(\mathcal{P}(S(\boldsymbol{y}) = \boldsymbol{s}_1 \mid A, y_0), \quad \ldots, \quad \mathcal{P}(S(\boldsymbol{y}) = \boldsymbol{s}_{2T} \mid A, y_0)\Big)'.$$

As done in (3.1) for our benchmark model, the second component in (3.7) can be reformulated as:

$$\boldsymbol{\mathcal{P}}^S = \int_{\mathcal{A}} \boldsymbol{\mathcal{L}}^S(A, \theta, y_0) dQ(A \mid y_0) = \tilde{G}(\theta) \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots A^{2T-1} \end{pmatrix} \frac{1}{\tilde{g}(A, \theta, y_0)} dQ(A \mid y_0),$$

where $\boldsymbol{\mathcal{P}}^S$ denotes the vector of length $2T$ with elements corresponding to the probabilities of $S(\boldsymbol{y})$ taking certain values on its support, and $\tilde{g}(A, \theta, y_0)$ denotes a polynomial function of A of degree $2T-1$. The crucial observation is that the matrix $\tilde{G}(\theta)$ is a square $2T \times 2T$ matrix of full rank. Therefore, the second component of (3.7) does not produce any moment equality conditions. All of the information about θ in this component must be in the form of moment inequality conditions. This result suggests that there does not exist an asymptotically more efficient estimator than the CMLE for this particular model.[15] The CMLE for this model is in fact semi-parametrically efficient. A formal proof is provided in the corrigendum by Gu, Hahn, and Kim (2021).

The above discussion of model (2.1) without covariates may leave the reader with the impression that the CMLE approach always factors the model in the appropriate way, and that focusing on the component of the likelihood conditional on the sufficient statistic will lead to an asymptotically efficient estimator. However, when covariates X are introduced, the only sufficient statistic is the original vector of choice histories such that $S(\boldsymbol{y}) = \boldsymbol{y}$, unless we impose a support restriction on X, as in Honoré and Kyriazidou (2000). Hence, the sufficient statistic of the fixed effects does not reduce the 2^T vector $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ to the lower dimensional subspace where it lives.[16] Fortunately, we can still apply our methodology to achieve the necessary reduction in dimension, if any.

4 Identification of Functionals of Unobserved Heterogeneity

We now turn to the situation where our parameters of interest are functionals of the latent distribution. Let ψ be some function and define the parameter of interest to be:

$$\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A, \theta_0, \boldsymbol{x})] = \int_{\mathcal{A}} \psi(A, \theta_0, \boldsymbol{x}) dQ_0(A|y_0, \boldsymbol{x}),$$

[15] The constructions of $\tilde{G}(\theta)$ and $\tilde{g}(A, \theta, y_0)$ are available upon request. Given our finding on the non-existence of moment equality conditions from $\boldsymbol{\mathcal{P}}^S$, we revisit Theorem 1 in Hahn (2001), and have found an error in the proof of Theorem 1 in Hahn (2001).

[16] This is likely due to the fact that when sufficiency argument is used, we are viewing the fixed effects as individual parameters, and each is allowed to have its own distribution. While here, we are accounting for the fact that we are modeling the fixed effects as random variables from a *common* distribution $Q(\cdots|y_0, \boldsymbol{x})$.

for some $\boldsymbol{x} \in \mathcal{X}$, where θ_0 is the true value of the structural parameter, and $Q_0(A|y_0, \boldsymbol{x})$ is the true latent distribution.

Definition 4.1. *The sharp identified set for* $\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A, \theta_0, \boldsymbol{x})]$ *is* $[\ell b(\boldsymbol{x}), ub(\boldsymbol{x})]$ *where:*

$$\ell b(\boldsymbol{x}) = \inf_{\{\theta,Q\}\in\mathcal{I}^*(y_0,\boldsymbol{x})} \int_{\mathcal{A}} \psi(A, \theta, \boldsymbol{x}) dQ(A),$$
$$ub(\boldsymbol{x}) = \sup_{\{\theta,Q\}\in\mathcal{I}^*(y_0,\boldsymbol{x})} \int_{\mathcal{A}} \psi(A, \theta, \boldsymbol{x}) dQ(A),$$

in which $\mathcal{I}^*(y_0, \boldsymbol{x})$ *is the identified set of* (θ, Q) *defined in Definition 2.1.*

This bounding problem is not always tractable because, although we have characterized the identified set Θ^* for θ using a finite set of moment conditions, the bounds in Definition 4.1 involve optimizing over a set of distributions, and this set can contain infinitely many elements.

However, as illustrated in Section 2.2, we can reduce the characterization of the identified set of the latent distribution to a finite vector of generalized moments, which are themselves specific functionals of Q. As soon as the functional of interest can be related to these generalized moments, we can profile out the distribution Q in the optimization problem defined in Definition 4.1. Additionally, as soon as the generalized moment is point identified, any functionals that can be written as an injective function of the generalized moments are also point identified. As we will show, a number of interesting functionals including the average marginal effect of the lagged choice arise as such a case.

Formulation (3.1) reveals that we have an equivalent definition of the set in (2.3):

$$\mathcal{Q}(\theta, y_0, \boldsymbol{x}) = \Big\{Q : \boldsymbol{r}(\theta, \boldsymbol{x}) = \int_{\mathcal{A}} \begin{pmatrix}1 & A & \cdots & A^{2T-1}\end{pmatrix}' \frac{1}{g(A, \theta, \boldsymbol{x}, y_0)} dQ\Big\}, \tag{4.1}$$

where $\boldsymbol{r}(\theta, \boldsymbol{x}) = H(\theta, \boldsymbol{x})\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ is the transformation of the population probability vector $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ defined in Theorem 3.1. We can use this definition to deduce the result below.

Theorem 4.1. *For each* $\boldsymbol{x} \in \mathcal{X}$*, and each value of* $\theta \in \Theta^*$*, all distributions* Q *whose generalized moments coincide with* $\mathbb{E}_Q[A^j/g(A, \theta, \boldsymbol{x}, y_0)]$*, for* $j = 1, 2, \ldots, 2T-1$*, are observationally equivalent in the sense that they generate the same population choice probability as in (2.3).*

Note that in the statement of the theorem, the index of the moment sequence starts with 1 instead of 0. This deserves some explanation. By definition, the function $g(A, \theta, \boldsymbol{x}, y_0)$ is a polynomial function of A of degree $2T-1$. Therefore, there is a linear relationship between a non-zero constant and $\{1/g, A/g, \cdots, A^{2T-1}/g\}$, for all $A \in \mathcal{A}$ (i.e. these are $2T$ linearly independent functions that reside in $2T-1$ dimensional vector space). The moments $\mathbb{E}_Q[A^j/g]$, $j=1,2,\ldots,2T-1$ determine the 'zero' moment $\mathbb{E}_Q[1/g]$. This result can also be seen from the fact that the choice probabilities sum to one: $1=\mathbf{1}'\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$. Indeed, this property implies $1=\mathbf{1}'\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}=\mathbf{1}'G(\theta,\boldsymbol{x})\mathbb{E}_Q[V(A)/g]$, where $V(A)=(1,A,\ldots,A^{2T-1})'$ and $\mathbf{1}'G(\theta,\boldsymbol{x})$ is a known $1\times 2T$ vector given θ.

For the results on functionals, we distinguish between two cases: The case in which the identified set Θ^* is a singleton (so that θ is point identified), and the case in which the identified set Θ^* has more than one element.

When θ is Point Identified

When $\Theta^*=\{\theta_0\}$, we know, by Definition 2.1, the true distribution $Q_0(A|y_0,\boldsymbol{x})$ is a member of $\mathcal{Q}(\theta_0,y_0,\boldsymbol{x})$, defined in (4.1). Therefore, $\boldsymbol{r}(\theta_0,\boldsymbol{x})=\int_{\mathcal{A}}\begin{pmatrix}1 & A & \cdots & A^{2T-1}\end{pmatrix}'\frac{1}{g(A,\theta_0,\boldsymbol{x},y_0)}dQ_0(A|y_0,\boldsymbol{x})$, equals $H(\theta_0,\boldsymbol{x})\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ by Theorem 3.1. Since θ_0 is point identified, $\boldsymbol{r}(\theta_0,\boldsymbol{x})$ is point identified.

Theorem 4.2. *For model (2.1), if θ is point identified, and the product $\psi(A,\theta_0,\boldsymbol{x})g(A,\theta_0,\boldsymbol{x},y_0)$ is a polynomial function of A with a degree that is no larger than $2T-1$ such that:*

$$\psi(A,\theta_0,\boldsymbol{x})g(A,\theta_0,\boldsymbol{x},y_0)=\sum_{j=0}^{2T-1}\eta_j(\theta_0,\boldsymbol{x})A^j,$$

for some vector $\boldsymbol{\eta}(\theta_0,\boldsymbol{x})=(\eta_0(\theta_0,\boldsymbol{x}),\eta_1(\theta_0,\boldsymbol{x}),\ldots,\eta_{2T-1}(\theta_0,\boldsymbol{x}))$, then $\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A,\theta_0,\boldsymbol{x})]$ is point identified and equal to $\boldsymbol{\eta}(\theta_0,\boldsymbol{x})'\boldsymbol{r}(\theta_0,\boldsymbol{x})$.

When θ is Partially Identified

When Θ^* consists of more than one distinct element, then, for each $\theta\in\Theta^*$, every probability measure in the set $\mathcal{Q}(\theta,y_0,\boldsymbol{x})$ produces the same first $2T-1$ generalized moments $\boldsymbol{r}(\theta,\boldsymbol{x})$ as $H(\theta,\boldsymbol{x})\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$. This feature leads to a characterization of sharp bounds for $\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A,\theta_0,\boldsymbol{x})]$.

Theorem 4.3. *For model (2.1), if Θ^* is a non-singleton set, and the product $\psi(A,\theta_0,\boldsymbol{x})g(A,\theta_0,\boldsymbol{x},y_0)$ is a polynomial function of A with a degree that is no larger than $2T-1$ such that:*

$$\psi(A,\theta_0,\boldsymbol{x})g(A,\theta_0,\boldsymbol{x},y_0)=\sum_{j=0}^{2T-1}\eta_j(\theta_0,\boldsymbol{x})A^j,$$

for some vector $\boldsymbol{\eta}(\theta_0,\boldsymbol{x})=(\eta_0(\theta_0,\boldsymbol{x}),\eta_1(\theta_0,\boldsymbol{x}),\ldots,\eta_{2T-1}(\theta_0,\boldsymbol{x}))$, then the sharp bounds for $\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A,\theta_0,\boldsymbol{x})]$ are given by $[\ell b(\boldsymbol{x}),ub(\boldsymbol{x})]$ where:

$$\ell b(\boldsymbol{x})=\inf_{\theta\in\Theta^*}\boldsymbol{\eta}(\theta,\boldsymbol{x})'\boldsymbol{r}(\theta,\boldsymbol{x}),$$

$$ub(\boldsymbol{x}) = \sup_{\theta \in \Theta^*} \boldsymbol{\eta}(\theta, \boldsymbol{x})' \boldsymbol{r}(\theta, \boldsymbol{x}).$$

Theorems 4.2 and 4.3 provide sufficient conditions under which a functional of $Q_0(A|y_0, \boldsymbol{x})$ can be point identified, or sharply bounded, without the need to optimize with respect to the distribution Q. In fact, for Theorem 4.2, no optimization is needed at all, while for Theorem 4.3, the optimization is with respect to the *finite* dimensional parameters, whose feasible regions are defined by a finite set of moment equality and inequality conditions.

4.1 Examples of Functionals

We now give three examples for which either Theorem 4.2 or 4.3 is applicable. In these examples, it becomes clear how to construct $\boldsymbol{\eta}(\theta, \boldsymbol{x})$.

4.1.1 Average Marginal Effects

For model (2.1) without covariates, let us define the "transition probability" conditional on A as:

$$\Pi_k(A, \beta) = \frac{A(\exp(\beta))^k}{1 + A(\exp(\beta))^k}, \quad k = \{0, 1\}.$$

The average marginal effect (AME) of the lagged choice is then defined as

$$AME_{y_0} = \int_{\mathcal{A}} \{\Pi_1(A, \beta_0) - \Pi_0(A, \beta_0)\} dQ_0(A|y_0).$$

For model (2.1) with covariates, the transition probability needs to be defined conditional on certain values of the covariates. For exposition let us consider the case with one covariate. For fixed values $\boldsymbol{x} \in \mathcal{X}$ and $\tilde{x} \in \mathbb{R}$, define:

$$\Pi_{k,\tilde{x},\boldsymbol{x}}(A, \theta) = \mathbb{P}(Y_{iT+1} = 1 | Y_{iT} = k, X_{T+1} = \tilde{x}, X_{1:T} = \boldsymbol{x}, A, \theta), \quad k = \{0, 1\}.$$

This quantity is interpreted as the transition probability from period T to period $T+1$, for an individual, whose fixed effect is A, and whose covariates coincide with $\{\boldsymbol{x}, \tilde{x}\} = \{x_1, \ldots, x_T, \tilde{x}\}$ from period 1 to $T+1$. We then define the AME of the lagged choice as:

$$AME_{\tilde{x},\boldsymbol{x}} = \int_{\mathcal{A}} \{\Pi_{1,\tilde{x},\boldsymbol{x}}(A, \theta_0) - \Pi_{0,\tilde{x},\boldsymbol{x}}(A, \theta_0)\} dQ_0(A|y_0, \boldsymbol{x}).$$

Proposition 4.1.

1. *For model (2.1) with $\gamma = 0$, $\Pi_k(A, \beta) g(A, \beta, y_0)$ is a polynomial function of A with a degree that is no larger than $2T - 1$, for each $k \in \{0, 1\}$.*

2. *For model (2.1) with covariates, if $\tilde{x} \in \{x_2, \ldots, x_T\}$, then, for $k = \{0, 1\}$, $\Pi_{k,\tilde{x},\boldsymbol{x}}(A, \theta)g(A, \theta, \boldsymbol{x}, y_0)$ is a polynomial function of A with a degree that is no larger than $2T - 1$.*

With the result in Proposition 4.1, we can now apply Theorems 4.2 and 4.3 and show that the aggregated transition probabilities, $\int_{\mathcal{A}} \Pi_k(A, \beta_0) dQ_0(A|y_0)$ and $\int_{\mathcal{A}} \Pi_{k,\tilde{x},\boldsymbol{x}}(A, \theta_0) dQ_0(A|y_0, \boldsymbol{x})$, are point identified, as long as $T \geq 3$, for each $k \in \{0, 1\}$. When $T = 2$, they can be sharply bounded. Since the average marginal effect is just the difference between these probabilities, the same conclusion applies. This result echoes those obtained in Aguirregabiria and Carro (2020), who are the first to point out that the AME of the lagged choice is point identified for the AR(1) dynamic panel logit model as long as $T \geq 3$. Because Aguirregabiria and Carro (2020) take a sequential identification approach to establish the identification results on AME, they require that the structural parameters are point identified. Our result extends to cases where the structural parameters are partially identified.

Example 1. *We first consider model (2.1) with $T = 2$ and no covariates. For this model, as discussed in Section 2.2, there is no moment equality condition and the inequalities define the sharp bounds for β. Appendix A.4 provides the analytical forms of these bounds. For the AME of the lagged choice, we have:*

$$(\Pi_1(A, \beta) - \Pi_0(A, \beta))g(A, \beta, y_0) = \boldsymbol{\eta}(\beta)' \begin{pmatrix} 1 & A & A^2 & A^3 \end{pmatrix}'$$

with $\boldsymbol{\eta}(\beta) = \begin{pmatrix} 0 & B-1 & B-1 & 0 \end{pmatrix}'$ for $y_0 = 0$. Applying Theorem 4.3, the sharp bounds for AME are given by:

$$\left[\inf_{\beta \in \Theta^*} \boldsymbol{\eta}(\beta)'\boldsymbol{r}(\beta), \sup_{\beta \in \Theta^*} \boldsymbol{\eta}(\beta)'\boldsymbol{r}(\beta)\right]$$

Given the form of $\boldsymbol{r}(\beta)$ in Section 2.2, we have $\boldsymbol{\eta}(\beta)'\boldsymbol{r}(\beta) = (B-1)p_1$ with $p_1 = \mathbb{P}(Y = \{1, 0\}|Y_0 = 0)$. With the analytical sharp bounds for β available, the sharp bounds for AME are just a simple linear transformation. These bounds are often quite narrow, suggesting that moment inequality conditions are informative about β as well as the AME. We illustrate these bounds in Figure 1.

Example 2. *Consider model (2.1) with $T = 3$ and one covariate, and let $y_0 = 0$. In this model θ is point identified. Without loss of generality let $\tilde{x} = x_3$, then $g(A, \theta_0, \boldsymbol{x}, y_0) = (1 + AB_0C_0^{x_2})(1 + AB_0C_0^{x_3})(1 + AC_0^{x_1})(1 + AC_0^{x_2})(1 + AC_0^{x_3})$ in which $B_0 = \exp(\beta_0)$ and $C_0 = \exp(\gamma_0)$. We have:*

$$\begin{aligned}
&(\Pi_{1,\tilde{x},\boldsymbol{x}}(A, \theta_0) - \Pi_{0,\tilde{x},\boldsymbol{x}}(A, \theta_0))g(A, \theta_0, \boldsymbol{x}, y_0) \\
&= (B_0 - 1) \begin{pmatrix} 1 & A & A^2 & A^3 & A^4 & A^5 \end{pmatrix} \boldsymbol{\eta}(\theta_0, \boldsymbol{x}, \tilde{x}),
\end{aligned}$$

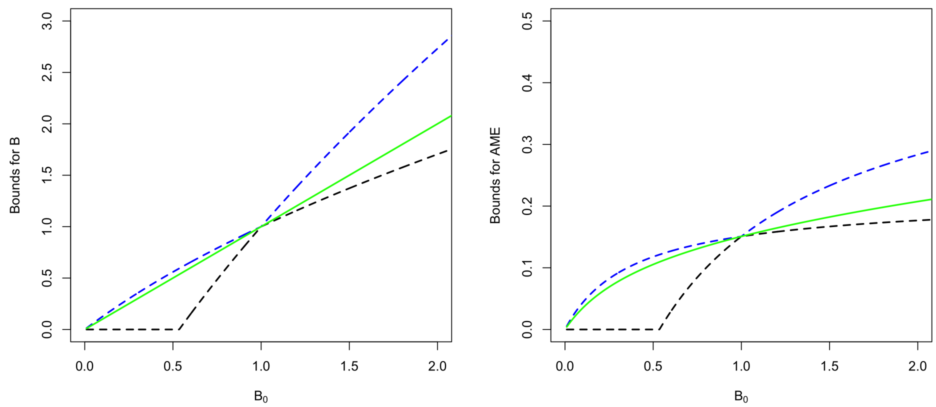

Figure 1: We illustrate the bounds for both the structural parameter as well as the average marginal effect for the model with two periods and no covariates as we vary B_0. For each values of β_0 ranging from log(0.01) to log(2), the data generating process assumes Q to be discrete with equal mass at -2 and 1 and $y_0 = 0$. Green solid line illustrates the true value B_0 and AME; blue dotted line illustrates the upper bound; black dotted line illustrates the lower bound.

where $\boldsymbol{\eta}(\theta_0, \boldsymbol{x}, \tilde{x}) = C_0^{x_3}\Big(0, \quad 1, \quad C_0^{x_1} + C_0^{x_2} + B_0 C_0^{x_2}, \quad C_0^{x_2}(C_0^{x_1} + B_0 C_0^{x_1} + B_0 C_0^{x_2}), \quad B_0 C_0^{x_1+2x_2}, \quad 0\Big)'$. *Applying Theorem 4.2, the average marginal effect is point identified as:*

$$AME_{\tilde{x},\boldsymbol{x}} = (\exp(\beta_0) - 1)\boldsymbol{\eta}(\theta_0, \boldsymbol{x}, \tilde{x})'\boldsymbol{r}(\theta_0, \boldsymbol{x}),$$

with $\boldsymbol{r}(\theta, \boldsymbol{x}) = H(\theta, \boldsymbol{x})\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$*, as defined in Theorem 3.1.*

In Section 5.1, we also consider model (2.1) with $T = 2$ and covariates, for which we derive the sharp bound of the AME in a specific numerical example. In Appendix A.5, similar results are presented for the case with $T = 3$ without covariates.

4.1.2 Posterior Mean of the Fixed Effects

In addition to the average marginal effect, researchers may also be interested in the posterior mean of a function of the fixed effects conditional on the observed choice history. This can be useful to infer the degree of heterogeneity across individuals, conditional on them making a certain sequence of choices. For simplicity, we focus on model (2.1) without covariates. For all $\boldsymbol{y}^j \in \mathcal{Y}$, define:

$$\varphi(A, \theta, \boldsymbol{y}^j) = A\mathcal{L}_j(A, \theta, y_0).$$

Proposition 4.2. *For model (2.1) with* $\gamma = 0$*, for all* $\boldsymbol{y}^j \in \mathcal{Y} \setminus \{1, 1, \ldots, 1\}$*,* $\varphi(A, \theta, \boldsymbol{y}^j)g(A, \theta, y_0)$ *is a polynomial function of* A *with a degree that is no larger than* $2T - 1$*.*

Applying Theorem 4.2 or 4.3, we can once again point identify, or construct sharp bounds for the following quantity:

$$\mathbb{E}[A|\boldsymbol{y}^j] = \frac{\int_{\mathcal{A}} A\mathcal{L}_j(A,\theta_0,y_0)dQ_0(A|y_0)}{\int_{\mathcal{A}} \mathcal{L}_j(A,\theta_0,y_0)dQ_0(A|y_0)} = \frac{\int_{\mathcal{A}} \varphi(A,\theta_0,\boldsymbol{y}^j)dQ_0(A|y_0)}{\mathcal{P}_j},$$

for all sequences of choices $\boldsymbol{y}^j \in \mathcal{Y} \setminus \{1,1,\ldots,1\}$. This quantity is the posterior mean of $\exp(\alpha)$ conditional on a particular choice history $\boldsymbol{y}^j$.

Example 3. *For model (2.1) with $\gamma = 0$, $y_0 = 0$ and $T = 3$. Consider the sequence of choices, $\boldsymbol{y} = \{0,1,0\}$, and let the observed choice probability be denoted by $\mathbb{P}_0(0,1,0)$, then we can show that:*

$$\mathbb{E}[\exp(\alpha)|\boldsymbol{y}] = \frac{1}{\mathbb{P}_0(0,1,0)} \begin{pmatrix} 0 & 0 & 1 & B_0+1 & B_0 & 0 \end{pmatrix} \boldsymbol{r}(\beta_0).$$

This conditional mean is point identified since $B_0 = \exp(\beta_0)$ and $\boldsymbol{r}(\beta_0) = H(\beta_0)\boldsymbol{\mathcal{P}}$ are both point identified given Theorem 3.1 and $\mathbb{P}_0(0,1,0)$ is directly identified from the data. More details on the form of $H(\beta_0)$ are provided in Appendix A.5.

4.1.3 Counterfactual Choice Probabilities with No Dynamics

Here we consider, as a counterfactual parameter of interest, the choice probability where there are no dynamics in the model, keeping everything else unchanged. This parameter contains information on how much of the choice persistence can be explained by the fixed effects, rather than the lagged choice dependence (i.e. $\beta \neq 0$).

For model (2.1) without covariates, we denote, for any sequence of choices $\boldsymbol{y}$,

$$\psi(A,\boldsymbol{y}) = \frac{A^{\sum_{t=1}^T y_t}}{(1+A)^T}, \tag{4.2}$$

and for the same model with one covariate, we denote:

$$\psi(A,\boldsymbol{y},\boldsymbol{x}) = \frac{A^{\sum_{t=1}^T y_t}\prod_{t=1}^T C^{x_t y_t}}{\prod_{t=1}^T (1+AC^{x_t})}. \tag{4.3}$$

Both quantities correspond to the counterfactual choice probability of $\boldsymbol{y} \in \mathcal{Y}$, conditional on the fixed effects taking value A. For the case with covariates, we further condition on the value of the covariates being $\boldsymbol{x} \in \mathcal{X}$.

Proposition 4.3.

1. *For model (2.1) without covariates, if $y_0 = 0$, then $\psi(A, \boldsymbol{y})g(A, \beta, y_0)$ is a polynomial function of A with a degree that is no larger than $2T - 1$.*

2. *For model (2.1) with covariates, if $y_0 = 0$, then $\psi(A, \boldsymbol{y}, \boldsymbol{x})g(A, \theta, \boldsymbol{x}, y_0)$ is a polynomial function of A with a degree that is no larger than $2T - 1$, for all $\boldsymbol{y} \in \mathcal{Y}$, and all $\boldsymbol{x} \in \mathcal{X}$.*

Applying Theorem 4.2 or 4.3, we can point identify, or construct sharp bounds for, the counterfactual choice probabilities, defined respectively by:

$$\begin{aligned}\text{no covariate:} \quad \mathcal{P}^\star(\boldsymbol{y}) &= \int_{\mathcal{A}} \psi(A, \boldsymbol{y}) dQ_0(A|y_0), \\ \text{with covariate:} \quad \mathcal{P}^\star_{\boldsymbol{x}}(\boldsymbol{y}) &= \int_{\mathcal{A}} \psi(A, \boldsymbol{y}, \boldsymbol{x}) dQ_0(A|y_0, \boldsymbol{x}),\end{aligned}$$

for any $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{x} \in \mathcal{X}$.

Example 4. *For model (2.1) with $T = 3$ and one covariate, we can point identify the structural parameters, and hence $B_0 = \exp(\beta_0)$ and $C_0 = \exp(\gamma_0)$ are known. For $y_0 = 0$, let us consider the choice history $\boldsymbol{y} = \{1, 1, 1\}$. It can be shown that:*

$$\psi(A, \boldsymbol{y}, \boldsymbol{x})g(A, \theta_0, \boldsymbol{x}, y_0) = A^3 C_0^{x_1+x_2+x_3} + A^4 B_0 C_0^{x_1+x_2+x_3}[C_0^{x_3} + C_0^{x_2} + B_0 C_0^{x_2+x_3}],$$

and, therefore, the counterfactual choice probability,

$$\mathcal{P}^\star_{\boldsymbol{x}}(1,1,1) = C_0^{x_1+x_2+x_3} \begin{pmatrix} 0 & 0 & 0 & 1 & B_0 C_0^{x_3} + C_0^{x_2} + B_0 C_0^{x_2+x_3} & 0 \end{pmatrix} \boldsymbol{r}(\theta_0, \boldsymbol{x}),$$

is point identified. Denote the observed conditional choice probability as $\mathbb{P}_{0,\boldsymbol{x}}(1,1,1)$, we can identify $\mathbb{P}_{0,\boldsymbol{x}}(1,1,1) - \mathcal{P}^\star_{\boldsymbol{x}}(1,1,1)$, which measures how much the state dependence contributes to the persistent choice.

5 Examples

In this section, we consider several additional examples. For models with very small T, we consider the two period model with a covariate. As special cases of the AR(1) models, we explicitly discuss the dynamic panel logit model with a time trend, as well as the model with time dummies. Last, we discuss extensions to the AR(2) model. Details of the AR(1) model with $T = 3$ or $T = 2$ without covariates are included in Appendices A.4 and A.5.

5.1 Two Periods with a Covariate

For simplicity, consider the case with one covariate and denote $C = \exp(\gamma)$. Maintain the assumption that $y_0 = 0$ and choose $g(A, \theta, \boldsymbol{x}, y_0) = (1 + AC^{x_1})(1 + AC^{x_2})(1 + ABC^{x_2})$ so that:[17]

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = (p_0(\boldsymbol{x}), p_1(\boldsymbol{x}), p_2(\boldsymbol{x}), p_3(\boldsymbol{x}))' = \int_{\mathcal{A}} G(\theta, \boldsymbol{x}) \begin{pmatrix} 1 & A & \cdots & A^3 \end{pmatrix}' \frac{1}{g(A, \theta, \boldsymbol{x}, y_0)} dQ(A \mid y_0, \boldsymbol{x}),$$

where $G(\theta, \boldsymbol{x})$ is defined by:

$$G(\theta, \boldsymbol{x}) = \begin{pmatrix} 1 & BC^{x_2} & 0 & 0 \\ 0 & C^{x_1} & C^{x_1+x_2} & 0 \\ 0 & C^{x_2} & BC^{2x_2} & 0 \\ 0 & 0 & BC^{x_1+x_2} & BC^{x_1+2x_2} \end{pmatrix}.$$

When $\beta \neq 0$, the matrix $G(\theta, \boldsymbol{x})$ is of full rank for any $\gamma \neq 0$, and any $\{x_1, x_2\} \in \mathbb{R}^2$. Therefore, the left null space of the matrix $G(\theta, \boldsymbol{x})$ is of zero dimension and there exist no moment equality conditions. In this setting, the transformed probability vector $\boldsymbol{r}(\theta, \boldsymbol{x})$ is:

$$\boldsymbol{r}(\theta, \boldsymbol{x}) = G^{-1}(\theta, \boldsymbol{x}) \boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = \frac{1}{B-1} \begin{pmatrix} (B-1)p_0 - B^2 C^{x_2 - x_1} p_1 + Bp_2 \\ BC^{-x_1} p_1 - C^{-x_2} p_2 \\ -C^{-(x_1+x_2)} p_1 + C^{-2x_2} p_2 \\ C^{-(x_1+2x_2)} p_1 - C^{-3x_2} p_2 + \frac{B-1}{B} C^{-(x_1+2x_2)} p_3 \end{pmatrix}, \tag{5.1}$$

for every $B, C \neq 1$. Applying Theorem 2.2 and Theorem 3.1, we get all the moment inequalities for each value of $\boldsymbol{x}$. The identified set for θ is the intersection of these inequalities, for all $\boldsymbol{x} \in \mathcal{X}$.

In Figure 2, we illustrate the bounds implied by this model given a specific choice for $\{\beta_0, \gamma_0\}$ and Q_0 under the assumption that $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\} = \{(1, 0), (0, 0)\}$. In this figure, we see that the sharp identified set Θ^* is rather small, even with only two possible values in the support $\mathcal{X}$.

Now we consider the bound for the average marginal effect in this model. Fix $\tilde{x} = 0$ and $y_0 = 0$, and consider the same data generating process used to obtain bounds in Figure 2. The average marginal effect can be specified as, for each $j \in \{1, 2\}$:

$$AME_{\tilde{x}, \boldsymbol{x}_j, y_0} = \int_{\mathcal{A}} \left\{ \frac{AB_0}{1 + AB_0} - \frac{A}{1 + A} \right\} dQ_0(A | y_0, \boldsymbol{x}_j).$$

[17] The elements in the set $\mathcal{Y}$ are ordered as: $\{(0,0), (1,0), (0,1), (1,1)\}$.

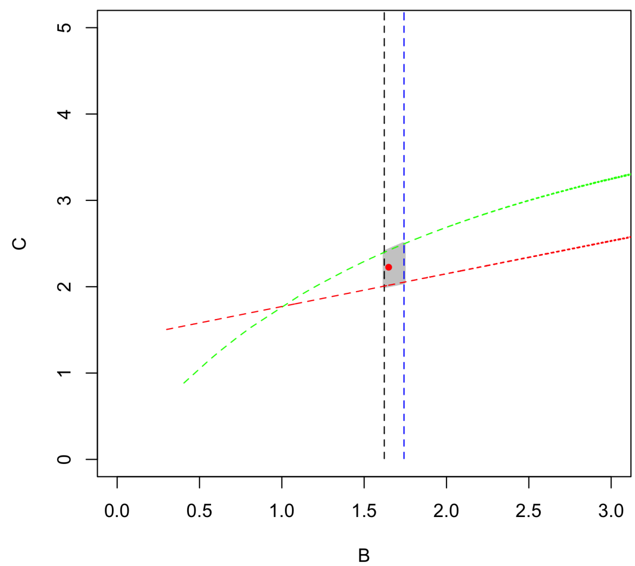

Figure 2: We illustrate the *binding* constraints imposed by the moment inequalities given $\mathcal{X} = \{(1,0),(0,0)\}$. For this figure, we assume that $Q_0(A|\boldsymbol{x}, y_0)$ is discrete with equal mass at -2 and 1 if $\boldsymbol{x} = (1,0)$ and is discrete with equal mass at -1 and -2 if $\boldsymbol{x} = (0,0)$, and that $(\beta_0, \gamma_0) = (0.50, 0.80)$. The shaded region is the sharp identified set; the red point illustrates the true parameters; blue dotted line illustrates the upper bound imposed by $r_0(\beta)r_2(\beta) - r_1(\beta)^2 \geq 0$ and the black dotted line illustrates the lower bound imposed by $r_1(\beta)r_3(\beta) - r_2(\beta)^2 \geq 0$ given $\boldsymbol{x} = (0,0)$; red dotted line corresponds to the lower bound imposed by $r_0(\theta)r_2(\theta) - r_1(\theta)^2 \geq 0$ and the green dotted line depicts the upper bound imposed by $r_1(\theta)r_3(\theta) - r_2(\theta)^2 \geq 0$ given $\boldsymbol{x} = (1,0)$. Constraints like $r_j \geq 0$ for all j are not binding and are not plotted for better visualization.

.

It is easy to verify that:

$$\Big\{\frac{AB}{1+AB} - \frac{A}{1+A}\Big\} g(A, \theta, \boldsymbol{x}_j, y_0) = \boldsymbol{\eta}(\theta, \boldsymbol{x}_j)' \begin{pmatrix} 1 & A & A^2 & A^3 \end{pmatrix}',$$

where $\boldsymbol{\eta}(\theta, \boldsymbol{x}_1)' = \begin{pmatrix} 0 & (B-1) & (B-1)C & 0 \end{pmatrix}$ and $\boldsymbol{\eta}(\theta, \boldsymbol{x}_2)' = \begin{pmatrix} 0 & (B-1) & (B-1) & 0 \end{pmatrix}$. Hence:

$$\begin{aligned} \boldsymbol{\eta}(\theta, \boldsymbol{x}_1)' \boldsymbol{r}(\theta, \boldsymbol{x}_1) &= \Big(\frac{B}{C} - 1\Big) p_1(\boldsymbol{x}_1) + (C-1) p_2(\boldsymbol{x}_1), \\ \boldsymbol{\eta}(\theta, \boldsymbol{x}_2)' \boldsymbol{r}(\theta, \boldsymbol{x}_2) &= (B-1) p_1(\boldsymbol{x}_2). \end{aligned}$$

Now we can apply Theorem 4.3 to bound the AME. In the particular data generating process that generates Figure 2, the true value of $AME_{\tilde{x}, \boldsymbol{x}_1, y_0}$ is 0.0749 with bounds $[0.0655, 0.0934]$, and the true value of $AME_{\tilde{x}, \boldsymbol{x}_2, y_0}$ is 0.0859, with bounds $[0.0828, 0.0979]$. Both sharp bounds are very informative and very simple to construct. The sharp bound for AME can be directly mapped from the identified set Θ^*.

5.2 Three Periods with a Time Trend

Now we consider the special case of the AR(1) model with three periods in which the only covariate is a time trend variable, hence the support of X, given by $\mathcal{X}$, contains elements of $\{1, 2, 3\}$. Under this specification, the likelihood function equals:

$$\mathcal{L}_j(A, \theta, y_0) = \prod_{t=1}^{3} \frac{\exp(\alpha + \beta y_{t-1} + \gamma t)^{y_t}}{1 + \exp(\alpha + \beta y_{t-1} + \gamma t)} = \frac{A^{\sum_{t=1}^{3} y_t} B^{\sum_{t=1}^{3} y_t y_{t-1}} C^{\sum_{t=1}^{3} t y_{it}}}{\prod_{t=1}^{3}(1 + AB^{y_{t-1}} C^t)},$$

where $A = \exp(\alpha)$, $B = \exp(\beta)$, and $C = \exp(\gamma)$. Let $\theta = \{\beta, \gamma\}$ and the matrix $G(\theta)$ is given by:

$$\begin{pmatrix}
1 & BC^2(1+C) & B^2C^5 & 0 & 0 & 0 \\
0 & C & C^3(1+BC) & BC^6 & 0 & 0 \\
0 & C^2 & C^4(B+C) & BC^7 & 0 & 0 \\
0 & C^3 & BC^5(1+C) & B^2C^8 & 0 & 0 \\
0 & 0 & BC^3 & BC^5(1+C) & BC^8 & 0 \\
0 & 0 & C^4 & C^6(1+BC) & BC^9 & 0 \\
0 & 0 & BC^5 & BC^7(B+C) & B^2C^{10} & 0 \\
0 & 0 & 0 & B^2C^6 & B^2C^8(1+C) & B^2C^{11}
\end{pmatrix},$$

with $g(A, \boldsymbol{x}, \theta, y_0) = (1 + AC)(1 + AC^2)(1 + AC^3)(1 + ABC^2)(1 + ABC^3)$ when $y_0 = 0$.

The left null space of the matrix $G(\theta)$ is spanned by the following two vectors:[18]

$$v_1 = \begin{pmatrix} 0 & C^3(B-1) & C(1-BC) & (C-1) & -BC^3(C-1) & BC^2(C-1) & 0 & 0 \end{pmatrix}'$$
$$v_2 = \begin{pmatrix} 0 & C^2(B-C) & C(1-B) & (C-1) & -BC^2(C-1) & 0 & (C-1) & 0 \end{pmatrix}'$$

It is easy to show that this basis cannot point identify (β, γ). The moment equalities implied by this basis have two non-trivial roots: one at the true parameters (β_0, γ_0), and one at a *false root* (see Figure 3 for an illustration). This model is also analyzed by Honoré and Weidner (2020). They lead to the same conclusion about false roots. This result implies that the structural parameters remain under-identified for a fixed value of y_0.[19] We show in this example how moment inequalities can help to rule out false roots. To illustrate, we consider a specific example. In particular, assume that Q_0 is discrete with equal mass at -2 and 1, and that $(\beta_0, \gamma_0) = (0.50, 0.80)$. The same DGP

[18] To construct this basis, we assume that $C \neq 1$.

[19] Honoré and Weidner (2020) propose to use the variation of y_0 to resolve this issue. It is true that we obtain two more moment conditions when $y_0 = 1$. Combining the four moment conditions allows one to point identify all structural parameters. However, our point here is that even if we fix $y_0 = 0$, we may already point identify the structural parameters by using the information from the moment inequalities. Later, we will also show an example (time dummy) where even by finding all moment equality conditions with variation in y_0, we still obtain multiple solutions to the system of moment equalities. But the information in the moment inequalities allows us to rule out false roots, and even yields point identification in some examples.

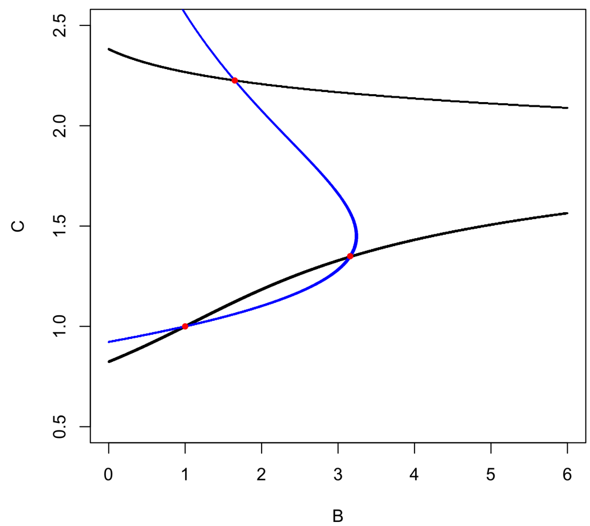

Figure 3: Black illustrates the curve on which the first moment equality holds; blue illustrates the curve on which the second moment equality holds. For this figure, we assume that Q is discrete with equal mass at -2 and 1, and that $(\beta_0, \gamma_0) = (0.50, 0.80)$. There are three solutions: the trivial root $B = C = 1$, the correct root, and the false root. Notice that the trivial root is assumed away in the construction of the moment equalities, leaving us with two roots.

is used to generate Figure 3. This specification yields the following population choice probabilities:

$$\boldsymbol{\mathcal{P}} \simeq \Big(0.0924, 0.0226, 0.0458, 0.1424, 0.0257, 0.0508, 0.1743, 0.4456\Big)'.$$

In this example, the moment equalities produce two non-trivial roots: one at the correct location $\theta_0 = (0.50, 0.80)$, and another roughly located at $\tilde{\theta} = (1.15, 0.30)$. We can rule out the false root by checking the non-negativity of the (Hankel) matrices in Theorem 2.2. In fact, in this particular example, it is sufficient for us to check only the sign of the second element of the transformed probability $\boldsymbol{r}(\theta)$. Indeed, by Theorem 2.2, it must be non-negative. Intuitively, it must be non-negative because it is the mean of a finite positive Borel measure. When we check these values, we find:

$$r_1(\theta_0) \simeq 0.01 \;\text{ and }\; r_1(\tilde{\theta}) \simeq -0.24,$$

where $r_1(\theta)$ denotes the second element of $\boldsymbol{r}(\theta)$. Therefore, ruling out the false root here is as simple as checking the sign of one transformed probability.

5.3 Three Periods with Time Dummies

We now consider a more complex example: the three-period dynamic panel logit model with time dummies. This model is characterized by $\gamma x_t = \gamma_t$, for $t = 1, 2, 3$. For simplicity, define $\gamma = \gamma_2$ and $\delta = \gamma_3$ (and $\gamma_1 = 0$ for normalization) and denote $\theta = \{\beta, \gamma, \delta\}$. Under this specification, the likelihood becomes:

$$\mathcal{L}_j(A, \theta, y_0) = \prod_{t=1}^{3} \frac{\exp(\alpha + \beta y_{t-1} + \gamma_t)^{y_t}}{1 + \exp(\alpha + \beta y_{t-1} + \gamma_t)} = \frac{A^{\sum_{t=1}^{3} y_t} B^{\sum_{t=1}^{3} y_t y_{t-1}} C^{y_2} D^{y_3}}{\prod_{t=1}^{3}(1 + A B^{y_{t-1}} C^{\mathbb{1}\{t=2\}} D^{\mathbb{1}\{t=3\}})},$$

where $A = \exp(\alpha)$, $B = \exp(\beta)$, $C = \exp(\gamma)$, and $D = \exp(\delta)$. The information content of this model was studied by Hahn (2001) and an earlier working paper version of Buchinsky, Hahn, and Kim (2010). They, however, considered only moment equality conditions from which they found zero information on β in the sense of Chamberlain (1992).[20] Using our approach, we are able to combine moment equality and inequality conditions to obtain sharp bounds for β. Our numerical illustration shows that the resulting sharp identified set can even be a singleton.

For this model we now have three parameters to consider. After integrating out the fixed effect, we obtain:[21]

$$\boldsymbol{\mathcal{P}} = (p_1, p_2, \ldots, p_8)' = \int_{\mathcal{A}} G(\theta) \begin{pmatrix} 1 & A & \cdots & A^5 \end{pmatrix}' \frac{1}{g(A, \theta, y_0)} dQ(A \mid y_0),$$

where the forms of $G(\theta)$ and $g(A, \theta, y_0)$ are included in Appendix A.7.

As before, we focus on $y_0 = 0$. The left null space of the matrix $G(\theta)$ is spanned by the following two vectors:

$$\begin{aligned} v_1 &= \begin{pmatrix} 0, & -BCD + BD^2, & -BCD, & -BCD, & C, & BD, & 0, & 0 \end{pmatrix}' \\ v_2 &= \begin{pmatrix} 0, & CD - D^2, & CD - BCD, & CD - D^2, & D - C/B, & 0, & -C + D, & 0 \end{pmatrix}'. \end{aligned} \tag{5.2}$$

Let us now characterize the set of all solutions to the moment equality conditions defined through (5.2), and then discuss whether there is any additional information contained in the moment inequalities. To start, we consider a linear combination of v_1 and v_2:

$$\frac{(v_1 + Bv_2)}{B} = \begin{pmatrix} 0, & 0, & -BCD, & -D^2, & D, & D, & -C + D, & 0 \end{pmatrix}'.$$

[20] We reproduce some of the results similar to Buchinsky, Hahn, and Kim (2010) in Appendix A.7, which have never been published, so to compare with our new results.

[21] The vector $\boldsymbol{\mathcal{P}}$ has elements $\mathbb{P}((Y_1, \ldots, Y_T) = \boldsymbol{y} | Y_0 = y_0)$ with $\boldsymbol{y} \in \mathcal{Y}$. The elements in the set $\mathcal{Y}$ are ordered as: $\{(1,1,1), (1,1,0), (1,0,1), (1,0,0), (0,1,1), (0,1,0), (0,0,1), (0,0,0)\}$.

The moment equality implied by the second vector above yields:

$$B = \frac{-D^2 p_4 + D(p_5 + p_6) + (-C + D)p_7}{CDp_3}. \tag{5.3}$$

Therefore, there exists a deterministic relationship between B and (C, D) given $\boldsymbol{\mathcal{P}}$. Consequently, the identification problem can be effectively reduced from a three parameter problem to a two parameter problem. The moment equality implied by the first vector remains. This moment equality, $v_1' \boldsymbol{\mathcal{P}} = 0$, can be written as:

$$\big\{(-CD + D^2)p_2 - CD(p_3 + p_4) + Dp_6\big\}\big\{-D^2 p_4 + D(p_5 + p_6) + (-C + D)p_7\big\} + C^2 D p_3 p_5 = 0. \tag{5.4}$$

This result implies that we can solve for the sharp identified set by finding all of the solutions (C^*, D^*) to (5.4), which is a polynomial functions of C and D. We can then use (5.3) to deduce B^* given each solution (C^*, D^*), and remove false solutions by checking the moment inequalities induced by $\boldsymbol{r}(\theta) \in \mathcal{M}_5$. Details on how to construct $\boldsymbol{r}(\theta)$ are included in the Appendix A.7.

We illustrate the power of moment inequalities in Figure 4. In the left panel, we see a curve in the positive orthant $\mathbb{R}^2_+$. This curve contains the solutions (C^*, D^*) to the moment equality in (5.4) under a specific choice of Q_0 and $(\beta_0, \gamma_0, \delta_0)$. We see that there are an uncountable number of non-trivial solutions from just the moment equality conditions. Interestingly and importantly, we find numerically that every false root is ruled out using moment inequalities.

We now investigate whether using variation in Y_0 will provide point identification for θ. As shown in Appendix A.7, for $y_0 = 1$, we find another moment equality condition that involves C and D, and again B can be written as a deterministic function of (C, D). However, since these moment equality conditions are polynomial functions of two variables, C and D, it is, in general, not clear how many real valued roots can be found from these two moment equalities. We can also not make a conclusive judgement on whether we can find a unique real valued solution. Indeed, using the Q_0 in the previous example, we find that the two moment equality conditions on (C, D) lead to *two* roots, one of them a false root. This is illustrated in the right panel of Figure 4. In this case, we just need to use our moment inequality conditions to check for these two candidate points, and then rule out the false root. In this example, the false root is $(B, C, D) \approx (1.646, 2.312, 2.308)$. Applying Theorem 2.2, we find that the second element of $\boldsymbol{r}(\theta)$ equals -0.179 for $y_0 = 0$, and -0.146 for $y_0 = 1$, at this parameter value, implying that this value of θ does not generate a vector $\boldsymbol{r}(\theta)$ that belongs to the moment space. Therefore, it is not in the identified set. In this example, combining the moment equality and inequality conditions allows us to reduce the sharp identified

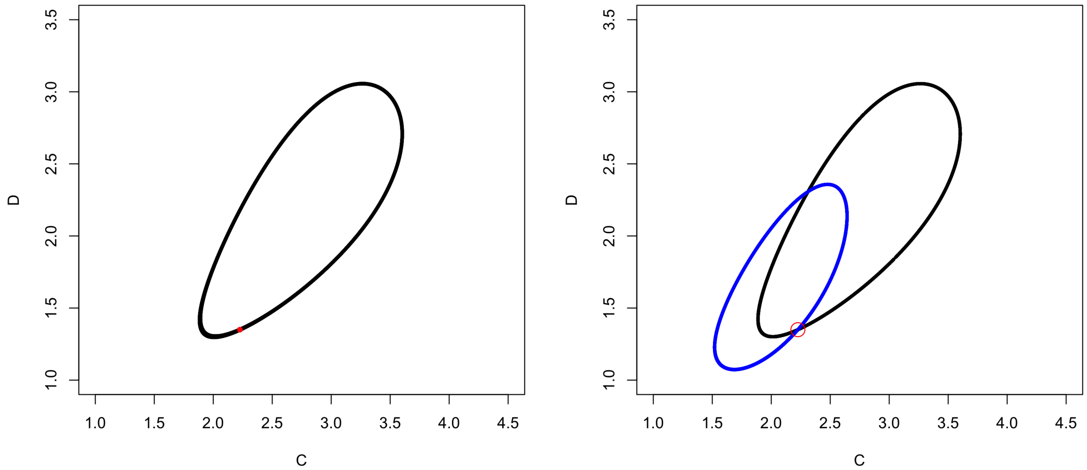

Figure 4: The left figure: Black illustrates the curve on which the moment equality in (5.4) holds, derived when $y_0 = 0$; the red point denotes the true solution. The right figure: the added blue curve illustrates the set of values of (C, D) on which the moment equality in (A.19) holds, derived when $y_0 = 1$ in Appendix A.7. The red circled point is again the true value. For both figures, we assume that Q_0 is discrete with equal mass at -2 and 1, and that $(\beta_0, \gamma_0, \delta_0) = (0.50, 0.80, 0.30)$.

set to a singleton.

5.4 AR(2) Model with Three Periods without Covariates

Now, we extend model (2.1) to the dynamic panel logit model with two lags, specified as:

$$Y_{it} = \mathbb{1}\{\alpha_i + \beta_1 Y_{it-1} + \beta_2 Y_{it-2} \geq \epsilon_{it}\}.$$

Assume that we observe $(Y_{-1}, Y_0, Y_1, Y_2, Y_3)$. For exposition, fix $(y_{-1}, y_0) = (0, 0)$. A similar analysis can be done with any $(y_{-1}, y_0) \in \{0, 1\}^2$. Variation in (Y_{-1}, Y_0) allows for an extra identifying constraint on the structural parameters since each value of $\{y_{-1}, y_0\}$ yields a set of moment conditions. Here, by fixing y_{-1} and y_0 at certain values, we consider the situation that the researchers do not have access to such variation.

Let $A = \exp(\alpha)$, $B_1 = \exp(\beta_1)$ and $B_2 = \exp(\beta_2)$. We can represent the likelihood of choice history (for general T) as:

$$\mathcal{L}_j(A, \theta, y_{-1}, y_0) = A^{\sum_{t=1}^T y_t} B_1^{\sum_{t=1}^T y_t y_{t-1}} B_2^{\sum_{t=1}^T y_t y_{t-2}} \sigma_A(0,0)^{m_1} \sigma_A(0,1)^{m_2} \sigma_A(1,0)^{m_3} \sigma_A(1,1)^{m_4}$$

with $\sigma_A(0,0) = \frac{1}{1+A}$, $\sigma_A(0,1) = \frac{1}{1+AB_1}$, $\sigma_A(1,0) = \frac{1}{1+AB_2}$ and $\sigma_A(1,1) = \frac{1}{1+AB_1B_2}$ and $m_1 = \sum_{t=1}^T (1 - y_{t-2})(1 - y_{t-1})$, $m_2 = \sum_{t=1}^T (1 - y_{t-2}) y_{t-1}$, $m_3 = \sum_{t=1}^T y_{t-2}(1 - y_{t-1})$ and $m_4 =$

$\sum_{t=1}^{T} y_{t-1}y_{t-2}$.

With $T=3$ and $\{y_{-1},y_0\}=\{0,0\}$, we have $m_1 \in \{1,2,3\}$ and $m_j \in \{0,1\}$, for $j=2,3,4$. By picking $g(A,\beta_1,\beta_2,y_{-1},y_0)=(1+A)^3(1+AB_1)(1+AB_2)(1+AB_1B_2)$, $\mathcal{L}_j(A,\theta,y_{-1},y_0)g(A,\beta_1,\beta_2,y_{-1},y_0)$ becomes a polynomial function of A with a degree no larger than 6, for all possible choice histories in the set $\mathcal{Y}$. Therefore, we again have the formulation:

$$\boldsymbol{\mathcal{P}} = G(\theta)\int_{\mathcal{A}}(1,A,\ldots,A^6)'\frac{1}{g(A,\theta,y_{-1},y_0)}dQ(A|y_{-1},y_0),$$

with $G(\theta)$ being a 8×7 matrix of full column rank provided $\beta_1,\beta_2\neq 0$. The particular form of $G(\theta)$ is available upon request.[22] The left null space of $G(\theta)$ has dimension equals to one, hence we expect one moment equality condition. Indeed, it takes the form:

$$-B_1\mathbb{P}_0(1,0,0)+B_1\mathbb{P}_0(0,1,0)-B_1\mathbb{P}_0(1,0,1)+\mathbb{P}_0(0,1,1)=0,$$

where $\mathbb{P}_0(y_1,y_2,y_3)$ denotes the choice probability of the choice history $(0,0,y_1,y_2,y_3)$. Clearly, β_1 is point identified from this moment equality, but β_2 is not. The identified set of θ can be constructed as:

$$\Theta^* = \{\theta=\{\beta_1,\beta_2\} : \exp(\beta_1)(-\mathbb{P}_0(1,0,0)+\mathbb{P}_0(0,1,0)-\mathbb{P}_0(1,0,1))+\mathbb{P}_0(0,1,1)=0, \boldsymbol{r}(\theta)\in\mathcal{M}_6\}.$$

Honoré and Weidner (2020) show that, as we vary the initial values $\{y_{-1},y_0\}$, we get more moment conditions, so that both β_1 and β_2 may be point identified. Our results do not contradict their results. When we have variation in the initial choices, we can take the intersection of all the implied restrictions on the structure parameters, including both the moment equalities and inequalities. For the case with $T=3$, we can point identify θ. However, our result becomes useful in situations where there is no or limited variation in $\{Y_{-1},Y_0\}$ in the population, so that the structural parameters are only partially identified.

We can also consider the identification of the average marginal effect of the lagged choices for this model. Denote the transition probability, conditional on A, as:

$$\Pi_{k_1,k_2}(A,\theta) := \mathbb{P}(Y_{t+1}=1|Y_t=k_1,Y_{t-1}=k_2,A,\theta)=\frac{AB_1^{k_1}B_2^{k_2}}{1+AB_1^{k_1}B_2^{k_2}}, \quad \{k_1,k_2\}\in\{0,1\}^2.$$

[22] We use a symbolic toolbox in Matlab to derive the form of G and H for all our examples. These code will be available on one of the authors' websites.

For any $\{k_1, k_2, \tilde{k}_1, \tilde{k}_2\} \in \{0,1\}^4$, the average marginal effect can be defined as:

$$AME_{y_{-1},y_0}(k_1, k_2, \tilde{k}_1, \tilde{k}_2) = \int_{\mathcal{A}} \Pi_{k_1,k_2}(A,\theta) - \Pi_{\tilde{k}_1,\tilde{k}_2}(A,\theta) dQ_0(A|y_{-1}, y_0).$$

It is easy to verify that $\Pi_{k_1,k_2}(A,\theta)g(A,\theta,y_{-1},y_0)$ is a polynomial function of A with a degree that is no larger than 6. For instance, consider $k_1 = k_2 = 0$, then $\Pi_{0,0}(A,\theta)g(A,\theta,y_{-1},y_0) = A(1+A)^2(1+AB_1)(1+AB_2)(1+AB_1B_2)$. By applying Theorem 4.2, we can then construct the sharp bound for the average marginal effect from the sharp identified set Θ^*.

5.5 AR(2) Model with Three Periods and a Covariate

Introducing one covariate in the AR(2) model leads to:

$$Y_{it} = \mathbb{1}\{\alpha_i + \beta_1 Y_{it-1} + \beta_2 Y_{it-2} + \gamma X_{it} \geq \epsilon_{it}\}.$$

We restrict attention to one covariate for ease of notation, but the framework easily extends to multiple regressors. For each value of $\boldsymbol{x} \in \mathcal{X}$, again denoting $A = \exp(\alpha)$, $B_1 = \exp(\beta_1)$, $B_2 = \exp(\beta_2)$ and $C = \exp(\gamma)$, and fixing $\{y_{-1}, y_0\} = \{0,0\}$, we have the following representation of the likelihood of a choice history $\boldsymbol{y}^j$:

$$\mathcal{L}_j(A,\theta,\boldsymbol{x},y_{-1},y_0) = A^{\sum_{t=1}^T y_t} B_1^{\sum_{t=1}^T y_t y_{t-1}} B_2^{\sum_{t=1}^T y_t y_{t-2}} C^{\sum_{t=1}^T y_t x_t} / \prod_{t=1}^{T} (1 + AB_1^{y_{t-1}} B_2^{y_{t-2}} C^{x_t}).$$

Take $g(A,\theta,\boldsymbol{x},y_{-1},y_0) = \prod_{t=1}^T (1+AC^{x_t}) \prod_{t=2}^T (1+AB_1C^{x_t}) \prod_{t=3}^T (1+AB_2C^{x_t})$, then it is easy to show that $\mathcal{L}_j(A,\theta,\boldsymbol{x},y_{-1},y_0)g(A,\theta,\boldsymbol{x},y_{-1},y_0)$ is a polynomial function of A, for all j. When $T = 3$, it is a polynomial function with a degree that is no larger than 7. Therefore, we again have the formulation:

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = G(\theta,\boldsymbol{x}) \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^7 \end{pmatrix}' \frac{1}{g(A,\theta,\boldsymbol{x},y_{-1},y_0)} dQ(A|\theta,\boldsymbol{x},y_{-1},y_0).$$

When $\theta \neq 0$ (all elements), the matrix $G(\theta,\boldsymbol{x})$ is of dimension 8×8 and has full rank. Therefore, the left null space is of zero dimension unless $x_2 = x_3$ as observed by Honoré and Weidner (2020). When $x_2 = x_3$, the rank of the matrix $G(\theta)$ is seven, and hence the left null space is of dimension one, yielding one moment equality for each distinct value of $\boldsymbol{x}$ such that $x_2 = x_3$. When it is not possible to impose the equality restriction on the covariate values, we can again consider the partial identification of the structural parameters using the moment inequalities due to Theorem 3.1. Our approach will be useful in this particular setting because there are no moment equalities unless

$x_2 = x_3$. This allows us to permit regressors that vary over time (i.e., time trend, age variable, or time dummies).

6 Estimation and Inference

While our main results concern identification, in this section, we demonstrate that some existing techniques from the literature can be used for estimation and inference. More specifically, we use results in Bajari, Benkard, and Levin (2007) and Shi and Shum (2015). One can develop a minimum distance framework that combines the moment equality and inequality conditions implied by Theorem 3.1. To illustrate, define a population objective function that combines identifying restrictions:

$$Q(\theta, \mathcal{P}) = m(\theta, \mathcal{P})' A' A m(\theta, \mathcal{P}) + \sum_{j=1}^{K} (\min\{h^j(\theta, \mathcal{P}), 0\})^2,$$

where θ denotes the structural parameters, $\mathcal{P}$ denotes the vector of (conditional) choice probabilities, $m(\theta, \mathcal{P})$ denotes a vector of equality conditions, $A'A$ denotes a nonsingular weighting matrix, and $h^j(\theta, \mathcal{P})$ $(j = 1, \ldots, K)$ denotes the inequality conditions. One may use all or only a subset of the moment inequalities implied by Theorem 3.1 (b).

Here the minimum distance setup is useful because it may facilitate the estimation procedure using our construction of inequality conditions, by imposing non-negativity on the two Hankel matrices, for which the generalized moments are written as a function of the choice probabilities. Without loss of generality, we assume that the inequality conditions hold such that $h^j(\theta, \mathcal{P}_0) \geq 0$ at the true parameter value θ_0, and hence the second term in $Q(\theta, \mathcal{P})$ becomes zero when θ satisfies the inequality conditions. Otherwise, it captures the departure from the inequality conditions when they are violated.

If the moment equality conditions alone can point identify the parameter (as in the AR(1) dynamic panel logit with $T \geq 3$), or all inequality conditions are slack so that they do not influence the asymptotic distribution of the estimator, one can instead use a GMM framework without using the inequality conditions at all. In this case, the estimation and inference can follow a standard GMM framework, but we consider a general setting for which both moment equality and inequality conditions are all potentially relevant.

Given this construction of the objective function, the true parameter vector θ_0 is a solution to:

$$\min_{\theta \in \Theta} Q(\theta, \mathcal{P}_0).$$

Our estimator then can be defined based on the sample analog of the criterion function, for the dynamic panel logit model without covariates, or with time trends or dummies, as:

$$Q_n(\theta, \hat{\mathcal{P}}_n) = m(\theta, \hat{\mathcal{P}}_n)' A_n' A_n m(\theta, \hat{\mathcal{P}}_n) + \sum_{j=1}^{K} (\min\{h^j(\theta, \hat{\mathcal{P}}_n), 0\})^2$$

or, for models with covariates x_i, as:

$$Q_n(\theta, \hat{\mathcal{P}}_n) = \frac{1}{n} \sum_i m(\theta, \hat{\mathcal{P}}_n(x_i))' A_n'(x_i) A_n(x_i) m(\theta, \hat{\mathcal{P}}_n(x_i)) + \frac{1}{n} \sum_i \sum_{j=1}^{K} (\min\{h^j(\theta, \hat{\mathcal{P}}_n(x_i)), 0\})^2,$$

and solves:

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} Q_n(\theta, \hat{\mathcal{P}}_n)$$

where $\hat{\mathcal{P}}_n$ denotes the vector of the estimated probabilities, which are sample frequencies or can be estimated using a sieve approach (e.g. Newey (1997), Chen (2007)) for models with covariates x_i.

For inference, we consider two cases. First, when θ_0 is point-identified, we can follow a standard inference method for extremum estimation. Second, when θ_0 is set-identified, one may derive the confidence bound using an existing method from the partial identification literature (e.g. Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2010), Andrews and Shi (2013), Pakes, Porter, Ho, and Ishii (2015) among others).

For the point-identified case, we assume the following conditions hold for all parameters in a small neighborhood of θ_0 (with probability approaching to one):

1. $B_0 = B(\theta_0)$ where $\sup_{\theta \in \Theta_0} \left\| \frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta, \hat{\mathcal{P}}_n) - B(\theta) \right\| = o_p(1)$
2. $\frac{\partial}{\partial\theta} Q_n(\theta, \hat{\mathcal{P}}_n) = \frac{\partial}{\partial\theta} Q_n(\theta_0, \hat{\mathcal{P}}_n) + (B_0 + o_p(1))(\theta - \theta_0)$
3. $\sqrt{n} \frac{\partial}{\partial\theta} Q_n(\theta_0, \hat{\mathcal{P}}_n) \to_d N(0, \Omega_0)$

We then have:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, B_0^{-1} \Omega_0 (B_0^{-1})').$$

These conditions can be verified straightforwardly under standard regularity conditions (see Newey and McFadden (1994)). For example, for the models such as the dynamic panel logit without covariates, or with time trend or dummy, following standard arguments, we get $\Omega_0 = \Lambda_0 \Sigma_0 \Lambda_0'$ where $\sup_{\mathcal{P} \in P_0} \left\| \frac{\partial^2}{\partial\theta\partial\mathcal{P}'} Q_n(\theta_0, \hat{\mathcal{P}}_n) - \Lambda_0 \right\| = o_p(1)$ and $\sqrt{n}(\hat{\mathcal{P}}_n - \mathcal{P}_0) \to_d N(0, \Sigma_0)$ which we utilize for our empirical application in Section 7.

For the partially identified case, we describe how one can use a subsampling method to construct confidence regions following Romano and Shaikh (2010), similar to Bajari, Benkard, and Levin (2007), among others. The resulting confidence set for the true parameter θ_0 should satisfy the coverage probability as

$$\liminf_{n\to\infty} \Pr(\theta \in \widehat{\Theta}_{(1-\alpha)}) \geq 1-\alpha$$

where $1-\alpha$ is a chosen confidence level. In the context of our minimum distance estimation, the subsample procedure is given as follows:

1. Choose a parameter set Θ such that it is certain that $\theta_0 \in \Theta$[23]
2. Construct B subsamples of size n_b and compute $Q_{n,b}(\theta, \hat{\mathcal{P}}_{n,b})$ for each subsample b and the parameter θ
3. For each $\theta \in \Theta$, compute a critical value $\hat{c}_n(1-\alpha, \theta)$ such that

$$\hat{c}_n(1-\alpha,\theta) = \inf\left\{c : \frac{1}{B}\sum_{b=1}^{B} 1\{n_b Q_{n,b}(\theta, \hat{\mathcal{P}}_{n,b}) \leq c\} \geq 1-\alpha\right\}$$

4. Compute $\widehat{\Theta}_{(1-\alpha)} = \left\{\theta : nQ_n(\theta, \hat{\mathcal{P}}_n) \leq \hat{c}_n(1-\alpha,\theta)\right\}$

There are several interesting estimation and inference questions that can be further explored. For instance, when only moment equality restrictions are considered, even though they are derived as conditional moment equalities, they can be converted into unconditional moment restrictions. For this case it is well known that, using the results from Chamberlain (1987), one can construct optimal instruments in the GMM framework to achieve the semiparametric efficiency bound. However, in our problem the identification restrictions can also include moment inequalities, which are in the form of conditional moment inequalities for models with covariates. In this case an efficiency result that combines both (conditional) moment equality and inequality restrictions are not developed yet to the best of our knowledge. Furthermore, when the support of $\mathcal{X}$ is rich or when we have continuous covariates, we may also face the usual curse of dimensionality. How one can utilize identification information from inequalities in an efficient way for estimation is an open question that we leave for future research.

[23] One can obtain a consistent estimator for the identified set, e.g. following Manski and Tamer (2002), as $\widehat{\Theta}_n = \left\{\theta : Q_n(\theta, \hat{\mathcal{P}}_n) \leq \min_{\theta\in\Theta} Q_n(\theta, \hat{\mathcal{P}}_n) + \kappa_n\right\}$ for some chosen $\kappa_n > 0$, such that $\kappa_n \to 0$.

7 Application

To demonstrate the use of our identification argument, we revisit the data analysis in Fitzmaurice and Laird (1993) on modeling children's respiratory conditions over a short period of time. The dataset, as part of the well known Harvard's Six Cities study conducted in the 1980s to study the association of air pollution and health outcomes, contains records on the wheezing condition (value 1 indicates yes, and value 0 indicates no) of 537 children from Steubenville, Ohio.[24] Each child is followed annually between the ages of 7 and 10. We also have information on whether the mother is a smoker in the first year of the study. We are interested in distinguishing between state dependence and the effect of age. It is plausible that respiratory diseases in young children may mitigate over age, in which case, we expect a downward time trend on the probability of having the wheezing condition for the sample of children. Not controlling for this time trend effect (or equivalently, the age effect) leads to biased estimates of the effect of state dependence that is, the persistence of wheezing. In addition, it is crucial to control for individual fixed effects in order to distinguish unobserved heterogeneity from the true dynamics. As the data also provides information on mother's smoking behaviour, we report estimation results for sub-samples depending on whether the mother is a smoker. Throughout the analysis, we focus on those children who have no wheezing condition in the initial period (at age 4), which consists of about 85% of the whole sample.

The model of interest is the same as in Section 5.2:

$$y_{it} = 1\{\alpha_i + \beta y_{it-1} + \gamma t \geq \epsilon_{it}\}, \quad t = 1, 2, 3$$

where ϵ_{it} are i.i.d random variables following the logistic distribution. We observe data $\{y_{i0}, y_{i1}, y_{i2}, y_{i3}\}$ for each child, and we are interested in the parameter (β, γ), while treating the α_i's as the incidental parameters. As illustrated in Section 5.2, there are two moment equality conditions for the parameter $(B, C) := (\exp(\beta), \exp(\gamma))$, and there are two solutions of (B, C) both satisfying the moment equality conditions. Denote $\hat{\mathcal{P}} = \frac{1}{n}\sum_{i=1}^{n} 1\{(y_{i1}, y_{i2}, y_{i3}) = \boldsymbol{y}\} := \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_8\}$, where the order of the choice history $\boldsymbol{y} = (y_1, y_2, y_3)$ is $\{(0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}$. The two moment equality conditions can then be written as[25]:

$$B = \frac{\hat{p}_7}{-C\hat{p}_3 + C^2(\hat{p}_1 + \hat{p}_6) + (C^3 - C^2)\hat{p}_5},$$

[24] The dataset is publically available in the R package `geepack`.

[25] We simplify the moment equality conditions presented in Section 5.2 and solve for B as a function of C. In particular, the first moment equality condition is derived from $(v_1 - v_2)'\hat{\mathcal{P}} = 0$ and the second is from $(Cv_1 - v_2)'\hat{\mathcal{P}} = 0$ where v_1 and v_2 are those presented in Section 5.2.

$$B = \frac{(C-1)\hat{p}_4 + C(\hat{p}_3 + \hat{p}_7) - C^3 \hat{p}_2}{C^2 \hat{p}_6}.$$

Figure 5 illustrates these conditions. The black line represents points of (B, C) that satisfy the first moment equality, while the red line represents the points of (B, C) that satisfy the second moment equality. The crossing points are the two roots that satisfy both moment equality conditions. Applying Theorem 3.1, as well as the estimation and inference method discussed in Section 6, we rule out one of the roots and get a unique solution combining information from moment equalities and inequality restrictions.

Results are reported in Table 1, labeled with *Logit Full*. The false root based on the full sample is $(\hat{\beta}, \hat{\gamma}) = (-0.088, -0.019)$. These estimates, while they satisfy the moment equality condition, they fail to satisfy the moment inequality conditions discussed in Theorem 3.1. In particular, the transformed probability $\boldsymbol{r}(\hat{\beta}, \hat{\gamma})$ based on the false root has its second entry equaling -4.82, which suggests that the vector $\boldsymbol{r}(\hat{\beta}, \hat{\gamma})$ cannot be a valid moment vector, since elements in the vector $\boldsymbol{r}$ are all supposed to be positive, representing moments of a non-negative measure supported on the positive real line. Checking non-negativity of the two Hankel matrices discussed in Theorem 2.2 confirms this finding.

As a comparison, we also report the estimates of the dynamic effect without controlling for a time trend. This reduces the model to a simple AR(1) model without covariates with $T = 3$. As we can see, we underestimate the persistence of the wheezing conditions substantially without controlling for time trend. The estimation results based on the sub-samples distinguishing whether a mother is a smoker or not presents a similar pattern, although they are estimated with more noise, as expected.

We also report results for several benchmark models, including the logit model without controlling for unobserved heterogeneity (labeled as *Logit*), as well as the logit fixed effect estimators (labeled as *Logit FE ML*), where all parameters, including the incidental parameters, are estimated with the maximum likelihood method. These estimates suggest that, without controlling for unobserved heterogeneity, the effects on persistence, and on the age effect, are over-estimated. For the *logit FE ML* estimates that control for individual effects, the incidental parameter problem leads to inconsistent estimates of (β, γ).

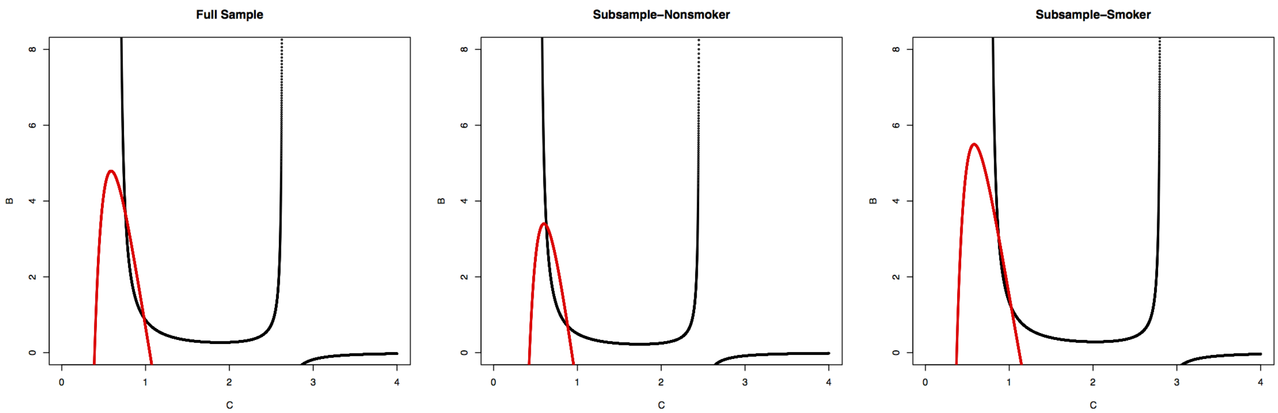

Figure 5: Illustration of multiple roots from the moment equality conditions of the AR(1) model with time trend. The left panel is for the whole sample, while the second and third panel is for sub-samples corresponding to mother as a smoker and a non-smoker respectively.

	Logit Full	*Logit Full*	*Logit*	*Logit*	*Logit FE ML*	*Logit FE ML*
			All Sample (n = 450)			
lagged y	1.301	0.693	2.08	1.772	-2.918	-2.736
	(0.671)	(0.707)	(0.258)	(0.238)	(0.690)	(0.503)
time trend	-0.276	-	1.05	-	1.666	-
	(0.321)	-	(0.162)	-	(0.260)	-
			Subsample: Smoker (n = 156)			
lagged y	1.210	0.693	2.189	1.809	-2.951	-2.889
	(1.937)	(1.225)	(0.400)	(0.359)	(1.17)	(0.856)
time trend	-0.464	-	1.082	-	1.654	-
	(0.338)	-	(0.254)	-	(0.412)	-
			Subsample: Non-smoker (n = 294)			
lagged y	1.118	0.693	1.972	1.708	-2.902	-2.649
	(1.101)	(0.866)	(0.342)	(0.320)	(0.855)	(0.623)
time trend	-0.135	-	1.038	-	1.674	-
	(0.513)	-	(0.211)	-	(0.335)	-

Table 1: Estimation results: *Logit Full* estimates the model using Theorem 3.1 and the estimation method discussed in Section 6. *Logit* estimates the model without individual fixed effects and *Logit FE ML* estimates all parameters in the model with the full likelihood, including the incidental parameters. Numbers in the brackets are standard errors.

8 Concluding Remarks

We characterize the sharp identified set for the structural parameters in a class of dynamic panel logit models. By reformulating the identification problem as a *truncated moment problem*, we show that all information on the structural parameters can be characterized by a set of moment equality and inequality conditions. We use this result to identify sharp bounds in models where structural parameters are not point identified, rule out false roots in models that cannot be identified using only moment equalities. We then characterize the observationally equivalent set of the latent distribution of fixed effects and show that we can only identify a finite vector of generalized moment of the latent distribution. Nevertheless, we provide conditions for a class of functionals of the latent distribution that can be point identified as soon as the structural parameters are point identified. We also discuss cases where functionals can be sharply bounded by only solving a simple finite dimensional optimization problem. We illustrate the usefulness of our results using a series of examples.

The connection to the truncated moment problem is due to the polynomial structure of the logit distribution with respect to the fixed effects. All panel logit models we consider enjoy this structure. Any other model beyond the logit model that has a similar polynomial structure can make use of our results. Our analytical approach to find moment equality conditions may be generalized to models with multi-dimensional fixed effects. These include the multinomial panel logit model and the bivariate models in which we consider choices of multiple interactive individuals (i.e., those considered in Honoré and Kyriazidou (2019), Honoré and de Paula (2021) and Aguirregabiria, Gu, and Mira (2021)). In these more complicated models, however, it can be challenging to generalize the results on the equivalence between model constraints on the generalized moment vector and a set of moment inequalities. This is due to the fact that the sum of square representation of non-negative polynomial functions only holds for the one-dimensional case. Nevertheless, the connection of the identification problem to the truncated moment problem may still be useful. We leave this for future research.

References

Aguirregabiria, V., and J. M. Carro (2020): "Identification of Average Marginal Effects in Fixed Effects Dynamic Discrete Choice Models," *Working Paper, University of Toronto.*

Aguirregabiria, V., J. Gu, and Y. Luo (2020): "Sufficient statistics for unobserved heterogeneity in structural dynamic logit models," *Journal of Econometrics, forthcoming.*

Aguirregabiria, V., J. Gu, and P. Mira (2021): "Identification of Structural Parameters in Dynamic Discrete Choice Games with Fixed Effects Unobserved Heterogeneity," *Working Paper.*

Andersen, E. B. (1970): "Asymptotic properties of conditional maximum-likelihood estimators," *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(2), 283–301.

Andrews, D., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

Arellano, M., and B. Honoré (2001): "Panel data models: some recent developments," in *Handbook of econometrics*, vol. 5, pp. 3229–3296. Elsevier.

Bajari, P., L. Benkard, and J. Levin (2007): "Estimating dynamic models of imperfect competition," *Econometrica*, 75(5), 1331–1370.

Bajari, P., J. Hahn, H. Hong, and G. Ridder (2011): "A note on semiparametric estimation of finite mixtures of discrete choice models with application to game theoretic models," *International Economic Review*, 52(3), 807–824.

Bonhomme, S. (2012): "Functional differencing," *Econometrica*, 80(4), 1337–1385.

Buchinsky, M., J. Hahn, and K. i. Kim (2010): "Semiparametric information bound of dynamic discrete choice models," *Economics Letters*, 108(2), 109–112.

Chamberlain, G. (1980): "Analysis of covariance with qualitative data," *The Review of Economic Studies*, 47(1), 225–238.

——— (1985): "Heterogeneity, duration dependence and omitted variable bias," *Longitudinal Analysis of Labor Market Data. Cambridge University Press New York.*

——— (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34(3), 305–334.

——— (1992): "Efficiency bounds for semiparametric regression," *Econometrica*, 60, 567–596.

Chebyshev, P. L. (1874): *Sur les valeurs limites des intégrales*. Imprimerie de Gauthier-Villars.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," *Handbook of Econometrics, Volume 6*, 6, 5549–5632.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and quantile effects in nonseparable panel models," *Econometrica*, 81(2), 535–580.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models," *Econometrica*, 75(5), 1243–1284.

Cressie, N., and P. W. Holland (1983): "Characterizing the manifest probabilities of latent trait models," *Psychometrika*, 48(1), 129–141.

Curto, R. E., and L. A. Fialkow (1991): "Recursiveness, positivity and truncated moment problems," *Houston Journal of Mathematics*, 17, 603–635.

Fitzmaurice, G. M., and N. M. Laird (1993): "A likelihood-based method for analysing longitudinal binary responses," *Biometrika*, 80(1), 141–151.

Gu, J., J. Hahn, and K. I. Kim (2021): "The information bound of a dynamic panel logit model with fixed effects – Corrigendum," *Econometric Theory*, p. 11.

Hahn, J. (2001): "The information bound of a dynamic panel logit model with fixed effects," *Econometric Theory*, 17, 913–932.

Heckman, J. J. (1981a): "Heterogeneity and state dependence," in *Studies in Labor Markets*, pp. 91–140. University of Chicago Press.

——— (1981b): "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," in *In Structural Analysis of Discrete Data with Econometric Applications*, pp. 179–195. Manski CF, McFadden D (eds). MIT Press: Cambridge, MA.

Honoré, B. E., and A. de Paula (2021): "Identification in Simple Binary Outcome Panel Data Models," *Cemmap Working Paper CWP14/21*.

Honoré, B. E., and E. Kyriazidou (2000): "Panel data discrete choice models with lagged dependent variables," *Econometrica*, 68(4), 839–874.

——— (2019): "Panel vector autoregressions with binary data," in *Panel Data Econometrics*, pp. 197–223. Elsevier.

Honoré, B. E., and E. Tamer (2006): "Bounds on parameters in panel dynamic discrete choice models," *Econometrica*, 74(3), 611–629.

Honoré, B. E., and M. Weidner (2020): "Moment Conditions for Dynamic Panel Logit Models with Fixed Effects," *ArXiv:2005.05942*.

Johnson, E. G. (2004): "Identification in discrete choice models with fixed effects," in *Working paper, Bureau of Labor Statistics*.

Karlin, S., and W. J. Studden (1966): *Tchebycheff systems: With applications in analysis and statistics*, vol. 15. Interscience Publishers.

Kitazawa, Y. (2021): "Transformations and moment conditions for dynamic fixed effects logit models," *Journal of Econometrics, forthcoming.*

Krein, M., and A. Nudelman (1977): "The Markov moment problem and extremal problems, Transl. Math," *Monographs, American Math. Soc., Providence*, 50.

Manski, C. F., and E. Tamer (2002): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70(2), 519–546.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79(1), 147–168.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Neyman, J., and E. L. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica*, 16, 1–32.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2015): "Moment Inequalities and Their Application," *Econometrica*, 83(1), 315–334.

Rasch, G. (1961): "On general laws and the meaning of measurement in psychology," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 4, pp. 321–333.

Romano, J. P., and A. M. Shaikh (2010): "Inference for the identified set in partially identified econometric models," *Econometrica*, 78(1), 169–211.

Shi, X., and M. Shum (2015): "Simple two-stage inference for a class of partially identified models," *Econometric Theory*, 31(3), 493–520.

Wooldridge, J. M. (2005a): "Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models," *Review of Economics and Statistics*, 87(2), 385–390.

——— (2005b): "Simple solutions to the initial conditions problem for dynamic nonlinear panel data models with unobserved heterogeneity," *Journal of Applied Econometrics*, 20, 39–54.

A Appendix

A.1 Proofs of Lemmas, Theorems, and Propositions

Proof of Theorem 3.1

It suffices to prove the theorem for a specific value of $\boldsymbol{x} \in \mathcal{X}$. To show necessity, we fix a pair $(\theta, Q) \in \mathcal{I}^*(y_0, \boldsymbol{x})$ defined in Definition 2.1 and we show that conditions (a) and (b) have to hold. In particular, we know from (3.1) that

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = G(\theta, \boldsymbol{x}) \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots A^{2T-1} \end{pmatrix}' \frac{1}{g(A, \theta, \boldsymbol{x}, y_0)} dQ(A)$$

with $g(A, \theta, \boldsymbol{x}, y_0)$ specified in Appendix A.2. Then it is easy to verify condition (a) by the definition of the set $\boldsymbol{M}_{\boldsymbol{x}}$. Condition (b) can be verified by the fact that $1/g(A, \theta, \boldsymbol{x}, y_0)$ is bounded on the support $\mathcal{A}$ as well as the fact that we can always construct $H(\theta, \boldsymbol{x}) = (G(\theta, \boldsymbol{x})' G(\theta, \boldsymbol{x}))^{-1} G(\theta, \boldsymbol{x})'$ such that we can find a finite positive Borel measure μ supported on $\mathcal{A}$ with $d\mu(A) = \frac{1}{g(A,\theta,\boldsymbol{x},y_0)} dQ(A)$ whose total mass and the first $2T-1$ moments are represented by $\boldsymbol{r}(\theta, \boldsymbol{x})$.

To show sufficiency, fix an arbitrary pair $(\theta, \boldsymbol{r}(\theta, \boldsymbol{x}))$ that satisfies the conditions (a) and (b) in Theorem 3.1. If there exists no moment equality condition, then this $(\theta, \boldsymbol{r}(\theta, \boldsymbol{x}))$ satisfies the condition (b) only. We will show we can always construct a probability measure Q supported on $\mathcal{A}$, given $(\theta, y_0, \boldsymbol{x})$, consistent with $\boldsymbol{r}(\theta, \boldsymbol{x})$, a vector of moments with respect to a Borel measure μ. In particular, we show that the constructed Q can generate the generalized moments, defined later, such that they are made identical to $\boldsymbol{r}(\theta, \boldsymbol{x})$. For this constructed Q, we also show the following (A.1) holds

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = \int_{\mathcal{A}} \boldsymbol{\mathcal{L}}(A, \theta, \boldsymbol{x}, y_0) dQ(A) = G(\theta, \boldsymbol{x}) \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^{2T-1} \end{pmatrix}' \frac{1}{g(A, \theta, \boldsymbol{x}, y_0)} dQ(A). \tag{A.1}$$

This implies $(\theta, Q) \in \mathcal{I}^*(y_0, \boldsymbol{x})$.

For ease of notation, below we suppress the dependence on these values $(\theta, y_0, \boldsymbol{x})$ in $Q, G, H, \boldsymbol{r}$ and g. First, note that the logit model (A.1) implies $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ has the following representation as $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = G \times \boldsymbol{c}$ where $\boldsymbol{c} \in \mathcal{M}_{2T-1}$, and hence $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ is a linear projection on G. We will show (1) $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = G \times \boldsymbol{r}$ if $\boldsymbol{r} \in \mathcal{M}_{2T-1}$, $\boldsymbol{r} = H\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$, and $HG = I$, and (2) we can construct a probability measure Q supported on $\mathcal{A}$ such that it generates a set of the generalized moments that match $\boldsymbol{r}$, and hence this Q and θ, satisfying the conditions (a) and (b), generate the model (A.1), concluding this (θ, Q) must be in $\mathcal{I}^*(y_0, \boldsymbol{x})$.

For (1), note that $\boldsymbol{r} = H\boldsymbol{\mathcal{P}_x} = HG \times \boldsymbol{c}$ for some $\boldsymbol{c} \in \mathcal{M}_{2T-1}$ since $\boldsymbol{\mathcal{P}_x}$ is a linear projection on G. It then follows that $\boldsymbol{r} = HG \times \boldsymbol{c} = \boldsymbol{c}$ since $HG = I$, and hence $\boldsymbol{\mathcal{P}_x} = G \times \boldsymbol{r}$.

For (2), let Q follow a discrete distribution supported on $2T$ distinct values, denoted as $\{a_1, \dots, a_{2T}\}$ with a probability measure $\pi_j = \mathbb{P}(A = a_j)$ such that $\sum_{j=1}^{2T} \pi_j = 1$. Note we are not fixing these values, any set of distinct support points work for our construction. Without loss of generality let $0 < a_1 < a_2 < \dots < a_{2T} < \infty$, then we show we can recover π_j's such that Q generates a set of generalized moments $\int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^{2T-1} \end{pmatrix}' / g(A) dQ$ such that they have identical values to the vector $\boldsymbol{r}$. We write a linear system of equations

$$A^g \overline{\pi} \equiv \begin{pmatrix} 1 & 1 & \cdots & 1 \\ a_1/g(a_1) & a_2/g(a_2) & \cdots & a_{2T}/g(a_{2T}) \\ \vdots & \vdots & \vdots & \vdots \\ a_1^{2T-1}/g(a_1) & a_2^{2T-1}/g(a_2) & \cdots & a_{2T}^{2T-1}/g(a_{2T}) \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{2T} \end{pmatrix} = \begin{pmatrix} 1 \\ r_1 \\ \dots \\ r_{2T-1} \end{pmatrix} \equiv \overline{\boldsymbol{r}}, \tag{A.2}$$

where r_j denotes the $j+1$-th element in the vector $\boldsymbol{r}$, the j-th moment with respect to a Borel measure μ. Here note that the zero-th moment $E_Q[1/g(A)]$ is redundant because the probabilities sum to one

$$1 = \mathbf{1}'\boldsymbol{\mathcal{P}_x} = \mathbf{1}'G \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^{2T-1} \end{pmatrix}' / g(A) dQ$$

for any Q, and in the system of equations we replace it with the condition $\sum_{j=1}^{2T} \pi_j = 1$, so that the resulting π_j's construct a proper distribution Q.

We know that A^g is nonsingular (a similar argument to Lemma 3.1), and we obtain $\overline{\pi} = (A^g)^{-1}\overline{\boldsymbol{r}}$. Finally, note that $1 = \mathbf{1}'\boldsymbol{\mathcal{P}_x} = \mathbf{1}'G \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^{2T-1} \end{pmatrix}' / g(A) dQ = \mathbf{1}'G\boldsymbol{r}$ implies $E_Q[1/g(A)] = r_0$ given Q matches the $2T-1$ generalized moments with $(r_1, \dots, r_{2T-1})$ in (A.2). This concludes (2).

Proof of Lemma 3.1

The set of functions in $\mathcal{V}_{\theta,\boldsymbol{x},y_0}(A)$ are linearly independent if $\sum_{j=1}^{2T} c_{j-1} A^{j-1} / g(A, \theta, \boldsymbol{x}, y_0) = 0$ holds only with constants $(c_0, \dots, c_{2T-1}) = 0$. Therefore, to prove the claim, it suffices to show that the determinant of the following $2T \times 2T$ matrix:

$$\begin{pmatrix} \frac{1}{g(a_0)} & \frac{1}{g(a_1)} & \cdots & \frac{1}{g(a_{2T-1})} \\ \frac{a_0}{g(a_0)} & \frac{a_1}{g(a_1)} & \cdots & \frac{a_{2T-1}}{g(a_{2T-1})} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{a_0^{2T-1}}{g(a_0)} & \frac{a_1^{2T-1}}{g(a_1)} & \cdots & \frac{a_{2T-1}^{2T-1}}{g(a_{2T-1})} \end{pmatrix}$$

where we write $g(a) = g(A = a, \theta, \boldsymbol{x}, y_0)$ given $(\theta, \boldsymbol{x}, y_0)$ for ease of notation, is non-zero for some distinct set of points $a_0, a_1, \ldots, a_{2T-1}$. Now take any distinct set of points such that $0 < a_0 < a_1 < \ldots < a_{2T-1} < \infty$, the determinant of the above matrix can be written as $\left(\prod_{j=0}^{2T-1} \frac{1}{g(a_j)}\right) \prod_{0 \le s < u \le 2T-1}(a_u - a_s)$, which is not equal to zero by construction.

Proof of Theorem 3.2

Given the representation (3.1) we have that $\boldsymbol{\mathcal{L}}(A, \theta, \boldsymbol{x}, y_0) = G(\theta, \boldsymbol{x})\mathcal{V}_{\theta,\boldsymbol{x},y_0}(A)$ for each $A \in \mathcal{A}$. The vectors in set $\kappa_{\boldsymbol{x}}$ satisfy that $\boldsymbol{m}'_{\boldsymbol{x}} G(\theta, \boldsymbol{x})$ is orthogonal to the vector $\mathcal{V}_{\theta,\boldsymbol{x},y_0}(A)$ for all $A \in \mathcal{A}$. Since the functions in $\mathcal{V}_{\theta,\boldsymbol{x},y_0}$ are linearly independent as shown in Lemma 3.1, the set of vectors $\mathcal{V}_{\theta,\boldsymbol{x},y_0}(A)$ spans $\mathbb{R}^{2T}$, hence the only vector of length $2T$ that can be orthogonal to $\mathbb{R}^{2T}$ is the null vector, i.e. $\boldsymbol{m}'_{\boldsymbol{x}} G(\theta, \boldsymbol{x}) = 0$, which precisely defined the set $\boldsymbol{M}_{\boldsymbol{x}}$.

Proof of Theorem 4.1

For ease of notation, given $(\theta, y_0, \boldsymbol{x})$, we write the choice probabilities, through (3.1), as

$$\boldsymbol{\mathcal{P}} = G \int_{\mathcal{A}} D(A) dQ(A) \equiv GD, \tag{A.3}$$

where Q is a probability distribution supported on $\mathcal{A}$, G is a $2^T \times 2T$ matrix, D is a $2T \times 1$ vector, and $D(A) = (1/g, A/g, \ldots, A^{2T-1}/g)'$ in our representation. In this case $\boldsymbol{\mathcal{P}}$ can be spanned by only (at most) $2T$ number of linearly independent vectors that span D since D is a $2T \times 1$ vector. We then have

$$\boldsymbol{\mathcal{P}} = G \sum_{l=1}^{2T} \pi_l D(a_l) \equiv G \int_{\mathcal{A}} D(A) d\tilde{Q}(A), \tag{A.4}$$

for some distinct values $\{a_1, \ldots, a_{2T}\}$ in the support $\mathcal{A}$ and weights $\{\pi_1, \ldots, \pi_{2T}\}$ that sums to 1 such that $D(a_l)$'s are linearly independent. Such a set exists because of Lemma 3.1. This implies for any given $Q(A)$ in (A.3) and a finite vector $\boldsymbol{\mathcal{P}}$ we can always construct an equivalent model to (A.3) using a finite mixture $\tilde{Q}$. Now note that, following a similar construction to the system of equations (A.2), once we know $2T-1$ generalized moments of $Q(A)$, we can find a discrete distribution $\tilde{Q}$ that satisfies (A.4). Therefore knowing $2T-1$ generalized moments of $Q(A)$ from (A.3) exhausts all information of $Q(A)$ we can learn from the model (A.3).

Proof of Theorem 4.2

The parameter of interest $\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A,\theta_0,\boldsymbol{x})]$ can be represented as

$$\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A,\theta_0,\boldsymbol{x})] = \sum_{j=0}^{2T-1} \eta_j(\theta_0,\boldsymbol{x})\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[A^j/g(A,\theta_0,\boldsymbol{x},y_0)].$$

Since θ is point identified, then we know θ_0. From Theorem 4.1, we know all measures in the set $\mathcal{Q}(\theta_0,y_0,\boldsymbol{x})$ defined in (4.1) have the same vector of generalized moments. Since $Q_0(A|y_0,\boldsymbol{x}) \in \mathcal{Q}(\theta_0,y_0,\boldsymbol{x})$ by construction, we have $r_j(\theta_0,\boldsymbol{x}) = \mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[A^j/g(A,\theta_0,\boldsymbol{x},y_0)]$, for $j=0,\ldots,2T-1$, which are observed quantities: given θ_0, $\boldsymbol{r}(\theta_0,\boldsymbol{x}) = H(\theta_0,\boldsymbol{x})\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$. Therefore $\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A,\theta_0,\boldsymbol{x})] = \boldsymbol{\eta}(\theta_0,\boldsymbol{x})'\boldsymbol{r}(\theta_0,\boldsymbol{x})$.

Proof of Theorem 4.3

We show the argument for sharp lower bound since the argument for the sharp upper bound is the same. By Definition 4.1, the lower bound of $\mathbb{E}_{Q_0(A|y_0,\boldsymbol{x})}[\psi(A,\theta_0,\boldsymbol{x})]$ can be written as

$$\begin{aligned}
\ell b(\boldsymbol{x}) &= \inf_{\theta\in\Theta^*, Q\in\mathcal{Q}(\theta,y_0,\boldsymbol{x})} \int_{\mathcal{A}} \psi(A,\theta,\boldsymbol{x})dQ(A) \\
&= \inf_{\theta\in\Theta^*, Q\in\mathcal{Q}(\theta,y_0,\boldsymbol{x})} \int_{\mathcal{A}} \boldsymbol{\eta}(\theta,\boldsymbol{x})' \begin{pmatrix} 1 & A & \ldots, A^{2T-1}\end{pmatrix}' \frac{1}{g(A,\theta,\boldsymbol{x},y_0)} dQ(A) \\
&= \inf_{\theta\in\Theta^*} \boldsymbol{\eta}(\theta,\boldsymbol{x})'\boldsymbol{r}(\theta,\boldsymbol{x}),
\end{aligned}$$

where the second equality is due to the fact that $\psi(A,\theta,\boldsymbol{x})g(A,\theta,\boldsymbol{x},y_0)$ can be represented as a polynomial funciton of A up to degree $2T-1$ with coefficients $\boldsymbol{\eta}(\theta,\boldsymbol{x})$. The last equality is due to the fact that all measures in the set $\mathcal{Q}(\theta,y_0,\boldsymbol{x})$, defined in (4.1), have the same vector of generalized moments, represented as $\boldsymbol{r}(\theta,\boldsymbol{x})$, which we observe for each given $\theta\in\Theta^*$.

Proof of Proposition 4.1

For the first result, using the form of $g(A,\beta,y_0)$ in Appendix A.2, for $k=0$, we have $\Pi_0(A,\beta)g(A,\beta,y_0) = A(1+A)^{T-1-y_0}(1+AB)^{T-1+y_0}$, which is a polynomial function of A of degree $2T-1$. For $k=1$, we have $\Pi_1(A,\beta)g(A,\beta,y_0) = AB(1+AB)^{T-2+y_0}(1+A)^{T-y_0}$, which is again a polynomial function of A of degree $2T-1$.

For the second result, we discuss four cases. If $k=0$ and $y_0=0$, then

$$\frac{AB^kC^{\tilde{x}}}{1+AB^kC^{\tilde{x}}}g(A,\theta,\boldsymbol{x},y_0)=\frac{AC^{\tilde{x}}}{(1+AC^{\tilde{x}})}\prod_{t=2}^{T}(1+ABC^{x_t})\prod_{t=1}^{T}(1+AC^{x_t}).$$

Since $\tilde{x}\in\{x_2,\ldots,x_T\}$, the right hand side is a polynomial function of A up to degree $2T-1$. If $k=0$ and $y_0=1$, then

$$\frac{AB^kC^{\tilde{x}}}{1+AB^kC^{\tilde{x}}}g(A,\theta,\boldsymbol{x},y_0)=\frac{AC^{\tilde{x}}}{(1+AC^{\tilde{x}})}\prod_{t=1}^{T}(1+ABC^{x_t})\prod_{t=2}^{T}(1+AC^{x_t}),$$

which is again a polynomial function of A up to degree $2T-1$. Similar argument applies for the case $k=1,y_0=0$ and $k=1,y_0=1$. We omit their forms for brevity.

Proof of Proposition 4.2

Note that

$$\mathcal{L}_j(A,\beta,y_0)=\frac{A^{n^{11}+n^{01}}B^{n^{11}}}{(1+AB)^{n^{11}+n^{10}}(1+A)^{n^{01}+n^{00}}},$$

with $n^{kj}=\sum_{t=1}^{T}1\{y_{t-1}=k,y_t=j\}$ for $k,j\in\{0,1\}$. Since $\max_{\boldsymbol{y}^j\in\mathcal{Y}\setminus\{1,\ldots,1\}}\left\{n^{11}+n^{01}\right\}\le T-1$ and the denominator of $\mathcal{L}_j(A,\beta,y_0)$ is always a polynomial function of A of degree T, then $\mathcal{L}_j(A,\beta,y_0)g(A,\beta,y_0)$ is polynomial of A up to degree $2T-2$, which implies that $A\mathcal{L}_j(A,\beta,y_0)g(A,\beta,y_0)$ is a polynomial of A up to degree $2T-1$.

Proof of Proposition 4.3

For the model (2.1) without covariates, given $g(A,\beta,y_0)=(1+AB)^{T-1+y_0}(1+A)^{T-y_0}$, we can verify

$$\psi(A,\boldsymbol{y})g(A,\beta,y_0)=A^{\sum_t y_t}(1+AB)^{T-1+y_0}(1+A)^{-y_0}.$$

Since $\sum_t y_t\in[0,T]$, when $y_0=0$, it is a polynomial function of A up to degree $2T-1$. For the model (2.1) with covariates, given $g(A,\theta,\boldsymbol{x},y_0)=\prod_{t=2-y_0}^{T}(1+ABC^{x_t})\prod_{t=1+y_0}^{T}(1+AC^{x_t})$, we can verify

$$\psi(A,\boldsymbol{y},\boldsymbol{x})g(A,\theta,\boldsymbol{x},y_0)=A^{\sum_{t=1}^{T}y_t}\prod_{t=1}^{T}C^{x_ty_t}\prod_{t=2-y_0}^{T}(1+ABC^{x_t})(1+AC^{x_1})^{-y_0}.$$

Then when $y_0=0$, it is a polynomial function of A up to degree $2T-1$ for any $\boldsymbol{x}\in\mathcal{X}$.

A.2 Generalized Moments Representation of AR(1) Model with General T

We first consider the dynamic logit model with T periods without covariates. The likelihood function for $\boldsymbol{y}^j = \{y_1, \ldots, y_T\}$ can be represented by

$$\mathcal{L}_j(A, \beta, y_0) = \frac{A^{n^{11}+n^{01}} B^{n^{11}}}{(1+AB)^{n^{11}+n^{10}}(1+A)^{n^{01}+n^{00}}},$$

with $n^{kj} = \sum_{t=1}^T 1\{y_{t-1} = k, y_t = j\}$ for $k, j \in \{0, 1\}$. Since $\max_{\boldsymbol{y}^j \in \mathcal{Y}}\{n^{11}+n^{10}\} = \max_{\boldsymbol{y}^j \in \mathcal{Y}} \sum_{t=1}^T 1\{y_{t-1} = 1\}$ and $\max_{\boldsymbol{y}^j \in \mathcal{Y}}\{n^{01}+n^{00}\} = \max_{\boldsymbol{y}^j \in \mathcal{Y}} \sum_{t=1}^T 1\{y_{t-1} = 0\}$, we can take $g(A, \beta, y_0) = (1+AB)^{T-1+y_0}(1+A)^{T-y_0}$. With this choice of $g(A, \beta, y_0)$, we can construct the matrix $G(\beta)$ of dimension $2^T \times 2T$ such that:

$$\boldsymbol{\mathcal{P}} = G(\beta) \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^{2T-1} \end{pmatrix}' \frac{1}{g(A, \beta, y_0)} dQ(A|y_0).$$

For model (2.1) with covariates, we allow the distribution of A to depend arbitrarily on the covariates, hence all the analysis is conditioned on $X = \boldsymbol{x} = \{x_1, x_2, \ldots, x_T\}$ (note we do not need to assume $\boldsymbol{x}$ to be a vector taking the same values, it can take any value in the support $\mathcal{X} \subset \mathbb{R}^T$). Here the derivation is made for one covariate. It can be easily generalized to multiple covariates with additional notation. The likelihood of $\boldsymbol{y}^j = \{y_1, y_2, \ldots, y_T\} \in \mathcal{Y}$ is then

$$\mathcal{L}_j(A, \theta, \boldsymbol{x}, y_0) = \frac{A^{\sum_{t=1}^T y_t} B^{\sum_{t=1}^T y_t y_{t-1}} C^{\sum_{t=1}^T x_t y_t}}{\prod_{t=1}^T (1 + AB^{y_{t-1}} C^{x_t})} = \frac{A^{\sum_{t=1}^T y_t} B^{\sum_{t=1}^T y_t y_{t-1}} C^{\sum_{t=1}^T x_t y_t}}{(1+ABC^{x_1})^{y_0}(1+AC^{x_1})^{1-y_0} \prod_{t=2}^T (1 + AB^{y_{t-1}} C^{x_t})}$$

with $B = \exp(\beta)$ and $C = \exp(\gamma)$. By taking $g(A, \theta, \boldsymbol{x}, y_0) = \prod_{t=2-y_0}^T (1 + ABC^{x_t}) \prod_{t=1+y_0}^T (1 + AC^{x_t})$, we can construct $G(\theta, \boldsymbol{x})$ of dimension $2^T \times 2T$ such that:

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = G(\theta, \boldsymbol{x}) \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^{2T-1} \end{pmatrix}' \frac{1}{g(A, \theta, \boldsymbol{x}, y_0)} dQ(A|\boldsymbol{x}, y_0).$$

A.3 Additional Discussion of Theorem 3.1

A.3.1 Alternative Construction of $H(\theta, x)$

In principle, the matrix $H(\theta, \boldsymbol{x})$ can always be constructed by $(G(\theta, \boldsymbol{x})' G(\theta, \boldsymbol{x}))^{-1} G(\theta, \boldsymbol{x})'$ to fulfill condition (b) in Theorem 3.1. However, sometimes it is more convenient to consider an alternative construction for $H(\theta, \boldsymbol{x})$ via the following procedure especially for $T > 2$.

Device a matrix H_0 of dimension $2T \times 2^T$ boolean matrix which picks $2T$ rows out of 2^T rows of the matrix $G(\theta, \boldsymbol{x})$, denote it as the reduced square matrix $\tilde{G}(\theta, \boldsymbol{x}) := H_0 G(\theta, \boldsymbol{x})$. Since $G(\theta, \boldsymbol{x})$ is full rank, so is the reduced $\tilde{G}(\theta, \boldsymbol{x})$. Perform LU factorization and write $\tilde{G}(\theta, \boldsymbol{x}) = LU$ and let $H(\theta, \boldsymbol{x}) = U^{-1}L^{-1}H_0$. This is more convenient because both U and L are upper or lower traingular matrices, which are easier to invert symbolically. Depending on the set of $2T$ rows one picks, there can exist multiple $H(\theta, \boldsymbol{x})$. That said, we do not need to check all possible choices of $H(\theta, \boldsymbol{x})$. To see this, given (3.1), we always have the relationship, $\boldsymbol{\mathcal{P}}_x = G(\theta, \boldsymbol{x})D(\theta, \boldsymbol{x})$, where $D(\theta, \boldsymbol{x}) = \int_{\mathcal{A}} \begin{pmatrix} 1 & A & \cdots & A^{2T-1} \end{pmatrix}' \frac{1}{g(A,\theta,\boldsymbol{x},y_0)} dQ(A|y_0, \boldsymbol{x})$ is the generalized vector of moments of A. Choosing any H such that $HG(\theta, \boldsymbol{x}) = I_{2T}$ leads to $\boldsymbol{r}(\theta, \boldsymbol{x}) = H\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = D(\theta, \boldsymbol{x})$. We may have multiple form of H that fulfills the condition $HG(\theta, \boldsymbol{x}) = I_{2T}$, which will just lead to different expression of $H\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ in terms of the elements in $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$. This is possible because some equations are redundant in the system $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}} = G(\theta, \boldsymbol{x})D(\theta, \boldsymbol{x})$.

A.3.2 Generalization to Situations Without the Full Rank Condition

For this discussion, we suppress the possible dependence of $G, \boldsymbol{\mathcal{P}}$, and g on $(\theta, y_0, \boldsymbol{x})$. If G is not of full column rank, then we can always write it as $G = G_0C$ where G_0 is comprised of linearly independent column vectors (i.e., a basis), and reformulate the choice probability equation (When G has full column rank we set C to be an identity matrix). Specifically, for the purpose of discussion, let G_0 be a 2^T by $2T - m$ matrix with $0 < m < 2T$ (so the rank of G equals to $2T - m$), and hence C is a $2T - m$ by $2T$ matrix. Let the vector of polynomials in A be $V(A) = (1, A, \ldots, A^{2T-1})'$ and the vector of the generalized moments be $V_g = \mathbb{E}_Q[V(A)/g(A)]$. Then the choice probability is given by $\boldsymbol{\mathcal{P}} = GV_g = G_0CV_g$ similar to (3.2), and we can rewrite this as $\boldsymbol{\mathcal{P}} = G_0\tilde{V}_g$ where $\tilde{V}_g = C\mathbb{E}_Q[V(A)/g(A)] = \mathbb{E}_Q[CV(A)/g(A)]$. By construction $CV(A)$ is another vector of polynomials in A.

Given this formulation, we can also decompose the degrees of freedom in $\boldsymbol{\mathcal{P}}$. The left null space of G_0 is of dimension $2^T - (2T - m)$ for any $T \geq 2$. Therefore, we can find $2^T - (2T - m)$ linearly independent vectors that form a basis for this space, where each vector serves as a moment equality condition for identifying θ. Note that this decomposition is now equal to the number of rows in $\boldsymbol{\mathcal{P}}$ (i.e., 2^T) minus the number of rows in $\tilde{V}_g$ (i.e., $2T - m$).

A.4 Details of the Model with Two Periods and No Covariates

For model (2.1) with $T = 2$ and $\gamma = 0$, we have

$$\boldsymbol{\mathcal{L}}(A,\beta,y_0) = \begin{pmatrix} \mathbb{P}((Y_1,Y_2)=(0,0)\mid Y_0=y_0,\alpha) \\ \mathbb{P}((Y_1,Y_2)=(1,0)\mid Y_0=y_0,\alpha) \\ \mathbb{P}((Y_1,Y_2)=(0,1)\mid Y_0=y_0,\alpha) \\ \mathbb{P}((Y_1,Y_2)=(1,1)\mid Y_0=y_0,\alpha) \end{pmatrix} = \begin{pmatrix} \frac{1}{(1+AB)^{y_0}(1+A)^{2-y_0}} \\ \frac{AB^{y_0}}{(1+AB)^{1+y_0}(1+A)^{1-y_0}} \\ \frac{A}{(1+AB)^{y_0}(1+A)^{2-y_0}} \\ \frac{A^2B^{y_0+1}}{(1+AB)^{1+y_0}(1+A)^{1-y_0}} \end{pmatrix}.$$

By choosing $g(A,\beta,y_0) = (1+A)^{2-y_0}(1+AB)^{1+y_0}$, we get:

$$\boldsymbol{\mathcal{L}}(A,\beta,y_0) = \begin{pmatrix} 1 & B & 0 & 0 \\ 0 & B^{y_0} & B^{y_0} & 0 \\ 0 & 1 & B & 0 \\ 0 & 0 & B^{y_0+1} & B^{y_0+1} \end{pmatrix} \begin{pmatrix} 1 \\ A \\ A^2 \\ A^3 \end{pmatrix} \frac{1}{g(A,\beta,y_0)}.$$

Now take the integral with respect to A and evaluate at $y_0 = 0$, and we get the expression in (2.4) as well as the form of the matrix $G(\beta)$.

Now, let $\boldsymbol{\mathcal{P}} = (p_0, p_1, p_2, p_3)'$, such that:

$$\boldsymbol{r}(\beta) = G(\beta)^{-1}\boldsymbol{\mathcal{P}} = \frac{1}{B-1}\begin{pmatrix} (B-1)p_0 - B^2p_1 + Bp_2 \\ Bp_1 - p_2 \\ -p_1 + p_2 \\ p_1 - p_2 + \frac{B-1}{B}p_3 \end{pmatrix}, \tag{A.5}$$

for every $B \neq 1$. By Theorem 3.1 and Theorem 2.2, a value of β is in the identified set Θ^* if and only if when evaluated at β, the following two (Hankel) matrices are non-negative:

$$H_1(\boldsymbol{r}(\beta)) = \begin{pmatrix} r_0(\beta) & r_1(\beta) \\ r_1(\beta) & r_2(\beta) \end{pmatrix} \text{ and } B_1(\boldsymbol{r}(\beta)) = \begin{pmatrix} r_1(\beta) & r_2(\beta) \\ r_2(\beta) & r_3(\beta) \end{pmatrix}$$

together with the range condition which states $\{r_2(\beta), r_3(\beta)\}$ is in Range$(H_1(\boldsymbol{r}(\beta)))$. The range condition deserves some extra explanation.

We distinguish two cases. The first case is when $H_1(\boldsymbol{r}(\beta))$ is singular (i.e., $\det(H_1(\boldsymbol{r}(\beta))) = 0$). In this case, there exists a constant $c_0 > 0$ such that $r_j(\beta) = c_0 r_{j-1}(\beta)$ for $j = 1, 2$. For $\{r_2(\beta), r_3(\beta)\}$ to be in Range$(H_1(\boldsymbol{r}(\beta)))$, we would have $r_3(\beta) = c_0 r_2(\beta)$. Since we can rule out $\boldsymbol{r}(\beta) = 0$ for any $\beta \neq 0$, then the non-negativity of $H_1(\boldsymbol{r}(\beta))$ and $B_1(\boldsymbol{r}(\beta))$ implies that

$r_0(\beta) > 0$. To summarize, the constraints on $\boldsymbol{r}(\beta)$ in this case include: $r_0(\beta) > 0$, $c_0 > 0$ and $r_j(\beta) = c_0 r_{j-1}(\beta)$ for $j = 1, 2, 3$. These conditions impose very strong restrictions on $\boldsymbol{r}(\beta)$. Yet, if $\boldsymbol{r}(\beta)$ is a moment sequence with $\det(H_1(\boldsymbol{r}(\beta))) = 0$, then we can only have a degenerate distribution Q since $\det(H_1(\boldsymbol{r}(\beta))) = r_0 r_2 - r_1^2$ is the variance of a probability measure.[26] If this variance is zero, then the measure must be degenerate.

The second case to consider is when $H_1(\boldsymbol{r}(\beta))$ is non-singular. In this case the range condition always holds, because we can solve the system of equations $H_1(\boldsymbol{r}(\beta))\boldsymbol{w} = (r_2, r_3)'$ and get a unique $\boldsymbol{w}$ which represents $(r_2, r_3)'$ as a linear combination of columns of $H_1(\boldsymbol{r}(\beta))$. It is also easy to show that in this case $r_j > 0$ for all $j = 0, 1, 2, 3$. To summarize, the conditions for $\boldsymbol{r}(\beta)$ to be a moment sequence include: $r_j > 0$, $r_0 r_2 - r_1^2 > 0$, $r_1 r_3 - r_2^2 \geq 0$. These constraints on $\boldsymbol{r}(\beta)$ map into constraints on the original vector $\boldsymbol{\mathcal{P}}$ which then lead to the following theorem.

Proposition A.1. *For model (2.1) with $T = 2$ and $\gamma = 0$, (i) the sign of β_0 is identified, and (ii) the following two cases define the sharp identified set Θ^*:*

1. *If $\beta_0 > 0$, then $\beta \in \Theta^*$ if, and only if:*
$$\frac{q_0 + \sqrt{q_0^2 - 4p_1p_2p_3(p_1 - p_2 + p_3)}}{2p_1(p_1 - p_2 + p_3)} \leq B \leq \frac{q_1 + \sqrt{q_1^2 + 4p_1p_2(p_0p_1 - p_0p_2 - p_2^2)}}{2p_1p_2}; \tag{A.6}$$
2. *If $\beta_0 < 0$, then $\beta \in \Theta^*$ if, and only if:*
$$\max\left\{0, \frac{q_1 - \sqrt{q_1^2 + 4p_1p_2(p_0p_1 - p_0p_2 - p_2^2)}}{2p_1p_2}\right\} \leq B \leq \frac{q_0 - \sqrt{q_0^2 - 4p_1p_2p_3(p_1 - p_2 + p_3)}}{2p_1(p_1 - p_2 + p_3)}; \tag{A.7}$$

where $B = \exp(\beta)$, $q_0 = p_1^2 - p_1p_2 + p_1p_3 + p_2p_3$ and $q_1 = p_0p_2 - p_0p_1 + p_1p_2 + p_2^2$.

The proof is given in the Appendix A.4.1.

A.4.1 Proof of Proposition A.1

We discuss the non-singular case when $\det(H_1(\boldsymbol{r}(\beta))) > 0$ and comment on the singular case where $\det(H_1(\boldsymbol{r}(\beta))) = 0$ at last.

For the non-singular case, as discussed in Section A.4, the identifying condition is $r_j > 0$ for $j = 0, \ldots, 3$, $r_0 r_2 - r_1^2 > 0$ and $r_1 r_3 - r_2^2 \geq 0$. The sign of β_0 is identified because $r_2 > 0$ if and only

[26]Recall $r_j(\beta) = \int_{\mathcal{A}} A^j d\bar{Q}(A, \beta)$ with $\bar{Q}(A, \beta)$ being some finite positive Borel measure and it can be made into a probability measure, denoted as $\tilde{Q}(A, \beta)$ by dividing $\bar{Q}(A, \beta)$ by r_0. Then consider $r_0 r_2 - r_1^2 = r_0^2(\frac{r_2}{r_0} - (\frac{r_1}{r_0})^2) = r_0^2(\int_{\mathcal{A}} A^2 d\tilde{Q}(A, \beta) - (\int_{\mathcal{A}} A d\tilde{Q}(A, \beta))^2)$. Since $r_0 > 0$ and the only possibility for $r_0 r_2 - r_1^2 = 0$ is $\tilde{Q}(A, \beta)$ having zero variance, which means $\tilde{Q}(A, \beta)$ is a degenerate probability measure.

if $B-1$ has the same sign as B_0-1. This result follows directly from the fact that $B_0>1$ implies $p_2>p_1$, and $B_0<1$ implies $p_2<p_1$. Since the sign of B_0-1 is identified, we can, without loss of generality, restrict our attention to two distinct cases: (a) B and B_0 larger than 1, and (b) B and B_0 smaller than 1.

Since in each case, B and B_0 have the same sign, the argument above implies $r_2>0$. Furthermore, the condition $r_1r_3-r_2^2\geq 0$ implies $r_1,r_3>0$. Similarly, $r_1>0$ and $r_0r_2-r_1^2>0$ implies $r_0>0$. Consequently, we only need to check $r_0r_2-r_1^2>0$ and $r_1r_3-r_2^2\geq 0$.

Consider case (a), B and B_0 are larger than 1. In this case $r_0r_2-r_1^2>0$ if and only if:

$$-B^2p_1p_2+B(p_0p_2-p_0p_1+p_1p_2+p_2^2)+(p_0p_1-p_0p_2-p_2^2)>0. \tag{A.8}$$

This expression is a quadratic equation with a discriminant equal to:

$$(p_0p_2-p_0p_1+p_1p_2+p_2^2)^2+4p_1p_2(p_0p_1-p_0p_2-p_2^2)>0.$$

This discriminant is strictly positive because $p_0,p_1,p_2\neq 0$ and $p_2\neq p_1$. Therefore, the quadratic equation in (A.8) has two distinct real-valued roots, and the quadratic formula implies that its roots have the form:

$$\frac{(p_0p_2-p_0p_1+p_1p_2+p_2^2)\pm\sqrt{(p_0p_2-p_0p_1+p_1p_2+p_2^2)^2+4p_1p_2(p_0p_1-p_0p_2-p_2^2)}}{2p_1p_2}. \tag{A.9}$$

Since the quadratic equation in (A.8) defines a parabola that opens down, the parameter B must be between these roots. Similarly, in this case, $r_1r_3-r_2^2\geq 0$ if and only if:

$$B^2p_1(p_1-p_2+p_3)-B(p_1^2-p_1p_2+p_1p_3+p_2p_3)+p_2p_3\geq 0. \tag{A.10}$$

The discriminant of this quadratic equation equals:

$$(p_1^2-p_1p_2+p_1p_3+p_2p_3)^2-4p_1p_2p_3(p_1-p_2+p_3)>0.$$

This discriminant is again strictly positive because $p_0,p_1,p_2\neq 0$ and $p_2\neq p_1$. Therefore, the quadratic equation in (A.10) has two distinct real-valued roots, and the quadratic formula implies that its roots have the form:

$$\frac{(p_1^2-p_1p_2+p_1p_3+p_2p_3)\pm\sqrt{(p_1^2-p_1p_2+p_1p_3+p_2p_3)^2-4p_1p_2p_3(p_1-p_2+p_3)}}{2p_1(p_1-p_2+p_3)}. \tag{A.11}$$

Since the quadratic equation in (A.10) defines a parabola that opens up, the parameter B cannot be between these roots. Finally, not all of these roots are needed: In particular, we can show that (i) the smaller root in (A.9) is equal to 1, (ii) the smaller root in (A.11) is smaller than 1, and (iii) the larger root in (A.11) is larger than 1. Together, these results imply the bounds in (A.6).

Define $C_0 = p_0(p_2 - p_1) + p_2(p_1 + p_2)$ and $D_0 = 4p_1p_2(p_0p_1 - p_0p_2 - p_2^2)$. To see that the smaller root in (A.9) is equal to 1, notice that, this root is equal to:

$$\frac{C_0 - \sqrt{C_0^2 + D_0}}{2p_1p_2},$$

and that this root is equal to 1 if and only if $C_0 - 2p_1p_2 = \sqrt{C_0^2 + D_0}$. Indeed, we can show that the left-hand side of this equality is strictly positive whenever $B_0 > 1$. Consequently, this root is equal to 1 if and only if:

$$-4p_1p_2C_0 + 4p_1^2p_2^2 = D_0.$$

It is easy to verify that this equality always holds by simply plugging in C_0 and D_0. Furthermore, it can be shown that $B_0 > 1$ implies $p_3 > p_2$, which, in turn, implies that the smaller root in (A.11) is smaller than 1. It is, therefore, left to show that the larger root in (A.11) is larger than 1. To see this result, let us define $C_1 = p_1D_1 + p_2p_3$, where $D_1 = p_1 - p_2 + p_3$, and assume that this root is, in fact, no larger than 1 such that:

$$\frac{C_1 + \sqrt{C_1^2 - 4p_1p_2p_3D_1}}{2p_1D_1} \leq 1.$$

This inequality leads to a contradiction because it holds if and only if:

$$\sqrt{C_1^2 - 4p_1p_2p_3D_1} \leq 2p_1D_1 - C_1 = p_1D + p_2p_3 = (p_1 + p_3)(p_1 - p_2) < 0.$$

Last we know the identified set is not empty because it has to contain the true points B_0. Therefore, we can rule out the case where the larger root of in (A.11) is located to the right of the larger root in (A.9), which will render an empty set for the identified set.

Consider case (b) where both B and B_0 are smaller than 1. In this case, $r_0r_2 - r_1^2 > 0$ if and only if

$$-B^2p_1p_2 + B(p_0p_2 - p_0p_1 + p_1p_2 + p_2^2) + (p_0p_1 - p_0p_2 - p_2^2) < 0. \tag{A.12}$$

and $r_1r_3 - r_2^2 \geq 0$ if and only if:

$$B^2p_1(p_1 - p_2 + p_3) - B(p_1^2 - p_1p_2 + p_1p_3 + p_2p_3) + p_2p_3 \leq 0. \tag{A.13}$$

We know that these quadratic equations have two distinct real-valued roots each, with the forms in (A.9) and (A.11). Because the quadratic equation in (A.12) defines a parabola that opens down, the parameter B cannot be between the roots in (A.9). Similarly, because the quadratic equation in (A.13) defines a parabola that opens up, the parameter B must be between the roots in (A.11).

Like before, not all of these roots are needed: It can be shown that $B_0 < 1$ implies $p_3 < p_2$, which, in turn, implies that the larger root in (A.11) is larger than 1. We can, therefore, ignore this root. Moreover, since $r_0 r_2 - r_1^2 > 0$ and $r_1 r_3 - r_2^2 \geq 0$ implies $B < 1$, it must be the case that these bounds, when considered together, yield an upper bound no larger than 1, implying that we can also ignore the larger root in (A.9).

Lastly for the singular case where $\det(H_1(\boldsymbol{r}(\beta))) = 0$, the condition for $B \in \Theta^*$ is that there exists $c_0 > 0$ such that $r_j = c_0 r_{j-1}$ for $j = 1, 2, 3$. This implies that $\det(B_1(\boldsymbol{r}(\beta))) = 0$. When $B_0 > 1$, there are only two possibilities for the identification condition to hold, either B equals to the larger root of (A.9) or the larger root of (A.11). Likewise, under the singular case if $B_0 < 1$, then there are only two possibilities for the identification condition to hold, either B equals to the smaller root of (A.9) or the smaller root of (A.11).

A.5 Details for AR(1) Logit Model with Three Periods and No Covariates

The choice probability for model (2.1) with $T = 3$ and $\gamma = 0$ can be written as

$$\boldsymbol{\mathcal{P}} = \int_{\mathcal{A}} G(\beta) \begin{pmatrix} 1 & A & \cdots & A^5 \end{pmatrix}' \frac{1}{g(A, \beta, y_0)} dQ(A|y_0),$$

with $G(\beta)$ taking the form:

$$G(\beta) = \begin{pmatrix} 1 & 2B & B^2 & 0 & 0 & 0 \\ 0 & 1 & 1+B & B & 0 & 0 \\ 0 & 1 & 1+B & B & 0 & 0 \\ 0 & 1 & 2B & B^2 & 0 & 0 \\ 0 & 0 & B & 2B & B & 0 \\ 0 & 0 & 1 & 1+B & B & 0 \\ 0 & 0 & B & B(1+B) & B^2 & 0 \\ 0 & 0 & 0 & B^2 & 2B^2 & B^2 \end{pmatrix},$$

and $g(A, \beta, y_0) = (1 + A)^3(1 + AB)^2$ for $y_0 = 0$.

The left null space of the matrix $G(\beta)$ is spanned by the following two vectors:[27]

$$v_1 = \begin{pmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}'$$
$$v_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -B & 1 & 0 \end{pmatrix}'$$

The moment equality condition implied by the second vector in this basis directly provides the point identification of β:

$$\beta_0 = \log\left(\mathbb{P}_0(0,1,1)\right) - \log\left(\mathbb{P}_0(1,0,1)\right).$$

To construct $\boldsymbol{r}(\beta)$, consider the matrix $H(\beta)$ to be

$$\frac{1}{(B-1)^2}\begin{pmatrix} (B-1)^2 & 0 & -B^2(2B-3) & B(B-2) & B^3 & -B^3 & 0 & 0 \\ 0 & 0 & B(B-2) & 1 & -B^2 & B^2 & 0 & 0 \\ 0 & 0 & 1 & -1 & B & -B & 0 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1/B & B-2 & 0 & 0 \\ 0 & 0 & -1 & 1 & \frac{B-2}{B} & 3-2B & 0 & \frac{(B-1)^2}{B^2} \end{pmatrix},$$

and we can verify $H(\beta)G(\beta) = I_6$.

Since β is point identified, we can apply Theorem 4.2 to show that the average marginal effect is point identified. The AME is defined as:[28]

$$\text{AME}_{y_0} = \int_{\mathcal{A}} \frac{AB_0}{1+AB_0} dQ_0(A|y_0) - \int_{\mathcal{A}} \frac{A}{1+A} dQ_0(A|y_0),$$

where $B_0 = \exp(\beta_0)$. When $y_0 = 0$, the AME can be written:

$$\text{AME}_0 = (B_0 - 1)\int_{\mathcal{A}} A(1+A)^2(1+AB_0)\frac{1}{g(A,\beta_0,y_0)} dQ(A \mid y_0).$$

There is a similar expression if $y_0 = 1$. We can verify

$$\begin{aligned} \text{AME}_0 &= (B_0 - 1)\int_{\mathcal{A}} A(1+A)^2(1+AB_0)\frac{1}{g(A,\beta_0,y_0)} dQ(A \mid y_0) \\ &= \begin{pmatrix} 0 & B_0 - 1 & (2+B_0)(B_0-1) & (1+2B_0)(B_0-1) & B_0(B_0-1) & 0 \end{pmatrix} \boldsymbol{r}(\beta_0) \\ &= (B_0 - 1)(\mathbb{P}_0(0,1,0) + \mathbb{P}_0(1,0,1)) \end{aligned}$$

Since β is point identified, the coefficients of this linear combination are known, and the AME

[27] To construct this basis, we assume that $B \neq 1$.

[28] We define the AME conditioning on Y_0. If the researcher wants to learn the AME without conditioning on Y_0 and if Y_0 is observed, then we may take an approach to integrate it out.

is point identified. Note that since there are multiple forms of $H(\beta)$ that satisfy the requirement $H(\beta)G(\beta) = I_6$, we may have different representation of AME_0. For example, we can also verify that $AME_0 = (B_0 - 1)\mathbb{P}_0(0,1,0) + \frac{B_0-1}{B_0}\mathbb{P}_0(0,1,1)$. But their values have to be the same if the model is correct, since we should have the same value for the generalized moments $\boldsymbol{r}(\beta)$ as long as $H(\beta)G(\beta) = I_6$.

A.6 Details for Section 3.2: AR(1) with Three Periods and a Covariate

For the example with three periods and one covariate in Section 3.2, the likelihood has the following form:

$$\mathcal{L}_j(A, \theta, \boldsymbol{x}, y_0) = \prod_{t=1}^{3} \frac{\exp(\alpha + \beta y_{t-1} + \gamma x_t)^{y_t}}{1 + \exp(\alpha + \beta y_{t-1} + \gamma x_t)} = \frac{A^{\sum_{t=1}^3 y_t} B^{\sum_{t=1}^3 y_t y_{t-1}} C^{\sum_{t=1}^3 x_t y_t}}{\prod_{t=1}^3 (1 + A B^{y_{t-1}} C^{x_t})},$$

where $A = \exp(\alpha)$, $B = \exp(\beta)$, and $C = \exp(\gamma)$.

When $y_0 = 0$, the matrix $G(\theta, \boldsymbol{x})$ has the following form:

$$\begin{pmatrix}
1 & B(C^{x_2} + C^{x_3}) & B^2 C^{x_2+x_3} & 0 & 0 & 0 \\
0 & C^{x_1} & C^{x_1}(C^{x_2} + BC^{x_3}) & BC^{x_1+x_2+x_3} & 0 & 0 \\
0 & C^{x_2} & C^{x_2}(BC^{x_2} + C^{x_3}) & BC^{2x_2+x_3} & 0 & 0 \\
0 & C^{x_3} & BC^{x_3}(C^{x_2} + C^{x_3}) & B^2 C^{x_2+2x_3} & 0 & 0 \\
0 & 0 & BC^{x_1+x_2} & BC^{x_1+x_2}(C^{x_2} + C^{x_3}) & BC^{x_1+2x_2+x_3} & 0 \\
0 & 0 & C^{x_1+x_3} & C^{x_1+x_3}(C^{x_2} + BC^{x_3}) & BC^{x_1+x_2+2x_3} & 0 \\
0 & 0 & BC^{x_2+x_3} & BC^{x_2+x_3}(BC^{x_2} + C^{x_3}) & B^2 C^{2x_2+2x_3} & 0 \\
0 & 0 & 0 & B^2 C^{x_1+x_2+x_3} & B^2 C^{x_1+x_2+x_3}(C^{x_2} + C^{x_3}) & B^2 C^{x_1+2x_2+2x_3}
\end{pmatrix}.$$

A.7 Details for Section 5.3

A.7.1 Information on β by Chamberlain (1992)

As in Section 5.3, we fix $y_0 = 0$. For the two basis vectors displayed in (5.2), we now show that there is no information on β using the Chamberlain (1992) approach. We begin by writing down the GMM representation induced by the moment equalities. Because of redundancy, we only need to consider the first 7 elements of $\mathcal{Y}$. Without loss of generality, let $\tilde{\mathbb{Y}}$ denote the first 7 elements of the vector $\mathbb{Y}$ with elements $\mathbb{1}\{(Y_1, \ldots, Y_T) = \boldsymbol{y}\}$ for $\boldsymbol{y} \in \mathcal{Y}$ and denote the corresponding probabilities as $\tilde{\boldsymbol{\mathcal{P}}}$. The moment restriction has the form:

$$\begin{aligned}
\mathbb{E}[v_1^{*\prime} \tilde{\mathbb{Y}} | Y_0 = y_0] &= 0 \qquad \text{(A.14)} \\
\mathbb{E}[v_2^{*\prime} \tilde{\mathbb{Y}} | Y_0 = y_0] &= 0,
\end{aligned}$$

where v_1^* and v_2^* denote the first 7 elements of v_1 and v_2 defined in (5.2), respectively. By Chamberlain (1992)' argument on semiparametric information, we know that the information for (B, C, D) implied by (A.14) is given by $\Delta'\Omega^{-1}\Delta$, where:

$$\Omega = \begin{pmatrix} v_1^{*\prime} \\ v_2^{*\prime} \end{pmatrix} \left(\text{diag}\left(\tilde{\boldsymbol{\mathcal{P}}}\right) - \tilde{\boldsymbol{\mathcal{P}}}\tilde{\boldsymbol{\mathcal{P}}}'\right) (v_1^* \; v_2^*) \quad \text{and} \quad \Delta = \begin{bmatrix} \tilde{\boldsymbol{\mathcal{P}}}'\boldsymbol{V}_1 \\ \tilde{\boldsymbol{\mathcal{P}}}'\boldsymbol{V}_2 \end{bmatrix},$$

in which we make use of the notation:

$$\boldsymbol{V}_1 = \begin{pmatrix} 0 & 0 & 0 \\ D(D-C) & -BD & B(2D-C) \\ -CD & -BD & -BC \\ -CD & -BD & -BC \\ 0 & 1 & 0 \\ D & 0 & B \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{V}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & D & C-2D \\ -CD & D-BD & C-BC \\ 0 & D & C-2D \\ \frac{1}{B^2}C & -\frac{1}{B} & 1 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}.$$

Now, notice that, Δ can be decomposed into $[\Delta_1, \Delta_2, \Delta_3]$, where:

$$\Delta_1 = \begin{pmatrix} -CD(p_2+p_3+p_4) + D(Dp_2+p_6) \\ -CDp_3 + \frac{1}{B^2}Cp_5 \end{pmatrix}, \quad \Delta_2 = \begin{pmatrix} p_5 - BD(p_2+p_3+p_4) \\ -p_7 + D(p_2+p_3+p_4-Bp_3) - \frac{1}{B}p_5 \end{pmatrix},$$

$$\text{and} \quad \Delta_3 = \begin{pmatrix} -BC(p_2+p_3+p_4) + B(2Dp_2+p_6) \\ p_5+p_7+C(p_2+p_3+p_4) - 2D(p_2+p_4) - BCp_3 \end{pmatrix}. \tag{A.15}$$

Therefore, with a little bit of algebra, it can be shown that, if we take the vector:

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = [\Delta_2 \; \Delta_3]^{-1}\,\Delta_1,$$

then we must obtain the following equality:

$$\Delta \begin{pmatrix} 1 & \lambda_1 & \lambda_2 \end{pmatrix}' = 0.$$

This result implies that the partial information for β contained in the moment equalities implied by (5.2) must be equal to zero, as the partial information for B implied by $\Delta'\Omega^{-1}\Delta$ is characterized by the minimum:

$$\min_{\lambda_1,\lambda_2} \begin{pmatrix} 1 & \lambda_1 & \lambda_2 \end{pmatrix} \Delta'\Omega^{-1}\Delta \begin{pmatrix} 1 & \lambda_1 & \lambda_2 \end{pmatrix}',$$

which must equal zero.

A.7.2 Detailed Construction for AR(1) Logit Model with the Time Dummy

In this part, we explicitly let results depend on y_0. For the time dummy case considered in Section 5.3, let

$$\boldsymbol{\mathcal{P}}^{y_0} = \begin{pmatrix} \mathbb{P}((Y_1,Y_2,Y_3)=(1,1,1)|Y_0=y_0) \\ \mathbb{P}((Y_1,Y_2,Y_3)=(1,1,0)|Y_0=y_0) \\ \mathbb{P}((Y_1,Y_2,Y_3)=(1,0,1)|Y_0=y_0) \\ \mathbb{P}((Y_1,Y_2,Y_3)=(1,0,0)|Y_0=y_0) \\ \mathbb{P}((Y_1,Y_2,Y_3)=(0,1,1)|Y_0=y_0) \\ \mathbb{P}((Y_1,Y_2,Y_3)=(0,1,0)|Y_0=y_0) \\ \mathbb{P}((Y_1,Y_2,Y_3)=(0,0,1)|Y_0=y_0) \\ \mathbb{P}((Y_1,Y_2,Y_3)=(0,0,0)|Y_0=y_0) \end{pmatrix} := \begin{pmatrix} p_1^{y_0} \\ p_2^{y_0} \\ p_3^{y_0} \\ p_4^{y_0} \\ p_5^{y_0} \\ p_6^{y_0} \\ p_7^{y_0} \\ p_8^{y_0} \end{pmatrix}.$$

When $y_0 = 0$, the matrix $G(\theta)$, now denoted as $G^0(\theta)$ to explicitly reflect its dependency on y_0, is defined by:

$$G^0(\theta) = \begin{pmatrix} 0 & 0 & 0 & B^2CD & B^2CD(C+D) & B^2C^2D^2 \\ 0 & 0 & BC & BC(C+D) & BC^2D & 0 \\ 0 & 0 & D & CD+BD^2 & BCD^2 & 0 \\ 0 & 1 & C+BD & BCD & 0 & 0 \\ 0 & 0 & BCD & BCD(BC+D) & B^2C^2D^2 & 0 \\ 0 & C & C(BC+D) & BC^2D & 0 & 0 \\ 0 & D & BD(C+D) & B^2CD^2 & 0 & 0 \\ 1 & B(C+D) & B^2CD & 0 & 0 & 0 \end{pmatrix},$$

and $g(A,\theta,y_0) = (1+A)(1+AC)(1+AD)(1+ABC)(1+ABD)$.

The vector $\boldsymbol{r}^0(\theta)$ in this case can be constructed from $H^0(\theta)\boldsymbol{\mathcal{P}}^0$ where $H^0(\theta)G^0(\theta) = I_6$. For instance consider $H^0(\theta)$ to be

$$\begin{pmatrix} 0 & \frac{B^2(C^2+D^2+CD)}{C(B-1)} & \frac{B^2CD-(C+D-BC)B^2(C+D)}{(B-1)(D-C)} & 0 & -\frac{(BD-C-D)B(C+D)+BCD}{D(B-1)(D-C)} & 0 & -\frac{B(C+D)}{D} & 1 \\ 0 & -\frac{B(C+D)}{C(B-1)} & \frac{BC+BD-B^2C}{(B-1)(D-C)} & 0 & \frac{BD-C-D}{D(B-1)(D-C)} & 0 & 1/D & 0 \\ 0 & \frac{1}{C(B-1)} & -\frac{1}{(B-1)(D-C)} & 0 & \frac{1}{BD(B-1)(D-C)} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{D(B-1)(D-C)} & 0 & -\frac{1}{BCD(B-1)(D-C)} & 0 & 0 & 0 \\ 0 & \frac{-1}{BC^2D(B-1)} & -\frac{1}{D^2(B-1)(D-C)} & 0 & \frac{1}{BC^2D(B-1)(D-C)} & 0 & 0 & 0 \\ \frac{1}{B^2C^2D^2} & \frac{D+C}{BC^3D^2(B-1)} & \frac{1}{D^3(B-1)(D-C)} & 0 & \frac{-1}{BC^3D(B-1)(D-C)} & 0 & 0 & 0 \end{pmatrix}.$$

The basis vectors of the left null space of $G^0(\theta)$ reduce to the moment condition for (C, D), as discussed in Section 5.3,

$$0 = (-CD + D^2)p_2^0 - CD(p_3^0 + p_4^0) + Dp_6^0 + \frac{C^2Dp_3^0p_5^0}{-D^2p_4^0 + D(p_5^0 + p_6^0) + (-C + D)p_7^0} \tag{A.16}$$

and B has a deterministic relationship with (C, D) as

$$B = \frac{-D^2p_4^0 + D(p_5^0 + p_6^0) + (-C + D)p_7^0}{CDp_3^0}. \tag{A.17}$$

Also, the moment inequality conditions are imposed through $\boldsymbol{r}^0(\theta) = H^0(\theta)\boldsymbol{\mathcal{P}}^0 \in \mathcal{M}_5$.

When $y_0 = 1$, we can make a similar derivation and have $G^1(\theta)$ as

$$\begin{pmatrix} 0 & 0 & 0 & B^3CD & B^3CD(C+D) & B^3C^2D^2 \\ 0 & 0 & B^2C & B^2C(C+D) & B^2C^2D & 0 \\ 0 & 0 & BD & BCD + B^2D^2 & B^2CD^2 & 0 \\ 0 & B & B(C+BD) & B^2CD & 0 & 0 \\ 0 & 0 & BCD & BCD(BC+D) & B^2C^2D^2 & 0 \\ 0 & C & C(BC+D) & BC^2D & 0 & 0 \\ 0 & D & BD(C+D) & B^2CD^2 & 0 & 0 \\ 1 & B(C+D) & B^2CD & 0 & 0 & 0 \end{pmatrix},$$

and the left null space of the matrix $G^1(\theta)$ is spanned by the following two vectors:

$$v_1 = \begin{pmatrix} 0 & -(C-D)/B & -C/B & -C/B & C/BD & 1 & 0 & 0 \end{pmatrix}'$$
$$v_2 = \begin{pmatrix} 0 & -D/B & -(CD-BCD)/B(C-D) & -D/B & (C-BD)/B(C-D) & 0 & 1 & 0 \end{pmatrix}'$$

or equivalently

$$v_1 = \begin{bmatrix} 0 \\ -(C-D)/B \\ -C/B \\ -C/B \\ C/BD \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and } -v_1 D + v_2(C-D) = \begin{bmatrix} 0 \\ 0 \\ CD \\ D^2/B \\ -D \\ -D \\ C-D \\ 0 \end{bmatrix}.$$

From $0 = Bv_1' \boldsymbol{\mathcal{P}}^1$ it follows that

$$B = \frac{(C-D)p_2^1 + C(p_3^1 + p_4^1) - Cp_5^1/D}{p_6^1}. \tag{A.18}$$

Also we obtain

$$\begin{aligned} 0 = (-v_1 D + v_2(C-D))' \mathcal{P}^1 &= CDp_3^1 + \frac{D^2 p_4^1}{B} - D(p_5^1 + p_6^1) + (C-D)p_7^1 \\ &= CDp_3^1 - D(p_5^1 + p_6^1) + (C-D)p_7^1 + \frac{D^3 p_4^1 p_6^1}{(C-D)Dp_2^1 + CD(p_3^1 + p_4^1) - Cp_5^1} \end{aligned} \tag{A.19}$$

where the third equality is obtained from (A.18).

Here to construct the vector of generalized moments $\boldsymbol{r}^1(\theta)$, we can take $H^1(\theta)$ as:

$$\begin{pmatrix} 0 & \frac{CBD-B(C+D)^2}{C(1-B)} & \frac{BCD-\{(C+D)-BC\}B(C+D)}{(B-1)(D-C)} & 0 & -\frac{\{BD-C-D\}B(C+D)+BCD}{D(B-1)(D-C)} & 0 & -\frac{B(C+D)}{D} & 1 \\ 0 & \frac{(C+D)}{C(1-B)} & \frac{(C+D)-BC}{(B-1)(D-C)} & 0 & \frac{BD-C-D}{D(B-1)(D-C)} & 0 & 1/D & 0 \\ 0 & \frac{-1}{BC(1-B)} & \frac{-1}{B(B-1)(D-C)} & 0 & \frac{1}{BD(B-1)(D-C)} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{BD(B-1)(D-C)} & 0 & -\frac{1}{BCD(B-1)(D-C)} & 0 & 0 & 0 \\ 0 & \frac{1}{B^2C^2D(1-B)} & -\frac{1}{BD^2(B-1)(D-C)} & 0 & \frac{1}{BC^2D(B-1)(D-C)} & 0 & 0 & 0 \\ \frac{1}{B^3C^2D^2} & \frac{C+D}{B^2C^3D^2(B-1)} & \frac{1}{BD^3(B-1)(D-C)} & 0 & \frac{1}{BC^3D(1-B)(D-C)} & 0 & 0 & 0 \end{pmatrix},$$

such that $H^1(\theta)G^1(\theta) = I_6$ and $\boldsymbol{r}^1(\theta) = H^1(\theta)\boldsymbol{\mathcal{P}}^1$. The moment inequality is imposed through $\boldsymbol{r}^1(\theta) \in \mathcal{M}_5$.