Transition Probabilities and Moment Restrictions in Dynamic Fixed Effects Logit Models *

Kevin Dano
Princeton University

October 6, 2024

Abstract

Dynamic logit models are popular tools in economics to measure state dependence. This paper introduces a new method to derive moment restrictions in a large class of such models with strictly exogenous regressors and fixed effects. We exploit the common structure of logit-type transition probabilities and elementary properties of rational fractions, to formulate a systematic procedure that scales naturally with model complexity (e.g the lag order or the number of observed time periods). We detail the construction of moment restrictions in binary response models of arbitrary lag order as well as first-order panel vector autoregressions and dynamic multinomial logit models. Identification of common parameters and average marginal effects is also discussed for the binary response case. Finally, we illustrate our results by studying the dynamics of drug consumption amongst young people inspired by Deza (2015).

Keywords: dynamic discrete choice, panel data, fixed effects.

JEL Classification Codes: C23, C33.

*I am very grateful to my advisors Bryan Graham, Stéphane Bonhomme, Demian Pouzo and Jim Powell for their generous support and advice. I also thank Chris Muris, Yassine Sbai Sassi, Bo Honoré, Martin Weidner, Adam Rosen, Stefan Hoderlein, Elena Manresa, Michal Kolesár, Ulrich Mueller, Bocar Ba, Nick Gebbia, Tahsin Saffat, seminar participants at UCSD, Emory, NYU, Zurich, LSE, Michigan, Duke, UPenn, Princeton, UCLA and the audiences at the 2023 IAAE Annual Conference, the 2023 California Econometrics Conference and the 2023 Causal Panel Data Conference at the Stanford Graduate School of Business for valuable comments and discussions. Financial support from the 2023 IAAE Conference is gratefully acknowledged. All errors are my own.

1 Introduction

The analysis of state dependence is a classic and important topic in many areas of economics. Several discrete processes such as welfare and labor force participation manifest strong serial persistence, and economists have sought various methods to unravel the underlying factors. In this paper, we reexamine the estimation of one notable set of models employed for this purpose: discrete choice models with lagged dependent variables, strictly exogenous regressors, fixed effects and logistic errors. We shall refer to this class of models as dynamic fixed effects logit models (DFEL) throughout. Specifications of this kind are used to discriminate between "structural" state dependence, i.e the causal effect of past choices on current outcomes, and heterogeneity, i.e the serial correlation induced by unobserved individual attributes (Heckman (1981)). An example of this approach is the analysis of welfare participation in Chay et al. (1999). There has been considerable interest in this family of panel data models in econometrics, with a recent surge in attention following new developments reported in Honoré and Weidner (2020). One general reason is that DFEL models stand out as a rare case of nonlinear dynamic panel data models for which solutions to the *incidental parameters problem* (Neyman and Scott (1948)) and *initial conditions problem* (e.g Heckman (1981)) have been known to exist in short panels[1].

In the "pure" version of the basic model which abstracts from covariates other than a first order lag, Cox (1958), Chamberlain (1985) and Magnac (2000) showed that the autoregressive parameter can be consistently estimated by conditional likelihood. This approach relies on the existence of a sufficient statistic linked to the logistic assumption to eliminate the fixed effect. In an important subsequent paper, Honoré and Kyriazidou (2000) extended this idea to a setting with strictly exogenous regressors and showed that the conditional likelihood approach remains viable if one can further condition on the regressors being equal in specific periods. This strategy was also found to be effective in dynamic multinomial logit models (Honoré and Kyriazidou (2000)), panel vector autoregressions (Honoré and Kyriazidou (2019)) and dynamic ordered logit models (Muris et al. (2020). At the same time, it has also

[1]The incidental parameters problem refers to the general inconsistency of maximum likelihood in short panels. The initial conditions problem refers to the general difficulty of formulating a correct conditional distribution for the initial observations given the fixed effects and covariates.

been noted that the necessity to be able to "match" the covariates imposes two limitations for the conditional likelihood approach: it inherently rules out time effects and implies rates of convergence slower than $\sqrt{N}$ for continuous explanatory variables. Furthermore, calculations from Honoré and Kyriazidou (2000) suggested that it does not easily extend to models with a higher lag order. These shortcomings have motivated the search for alternative methods of estimation.

Recently, Kitazawa et al. (2013, 2016) and Kitazawa (2022) revisited the AR(1) logit model - autoregressive of order one - of Honoré and Kyriazidou (2000) and proposed a transformation approach that deals with the fixed effects without restricting the nature of the covariates besides the conventional assumption of strict exogeneity. Their methodology leads to moment restrictions that can serve as a basis to estimate the model parameters at $\sqrt{N}$-rate by GMM; even with continuous regressors. In parallel work, Honoré and Weidner (2020) also derived moment conditions for the AR(1), AR(2) and AR(3) logit models in panels of specific length using the functional differencing technique of Bonhomme (2012). Their approach is partly numerical and relies on symbolic computing (e.g Mathematica) to obtain analytical expressions but has a wider scope of potential applications, e.g dynamic ordered logit specifications (Honoré et al. (2021)). In another recent paper, Dobronyi et al. (2021), the authors analyze the full likelihood of AR(1) and AR(2) logit models with discrete covariates under a new angle that reveals a connection to the *truncated moment problem* in mathematics. Drawing on well established results in that literature, they derive moment equality and new moment inequality restrictions that fully characterize the sharp identified set.

In this paper, we introduce a new systematic approach to construct moment restrictions in DFEL models with additive fixed effects, i.e when fixed effects are heterogeneous "intercepts". This class of models encompasses most specifications studied in prior work but excludes models with heterogeneous coefficients on lagged outcomes and/or regressors as in Chamberlain (1985) and Browning and Carro (2014). Unlike some recent competing approaches, we do not require numerical experimentation nor symbolic computing. Rather, as we shall see in examples, we exploit the common structure of logit-type transition probabilities and elementary properties of rational fractions, to obtain analytic expressions for the identifying moments. We shall focus our attention on deriving valid moment functions for

AR(p) models with arbitrary lag order $p \geq 1$ as well as first-order panel vector autoregressions and dynamic multinomial logit models (Magnac (2000)).

Our methodology exploits two key observations. First, the transition probabilities of logit-type models can often be expressed as conditional expectations of functions of observables and common parameters given the initial condition, the regressors and the fixed effects. We shall refer to these moment functions as *transition functions*. They have the important feature of not depending on individual fixed effects. Second, as soon as $T \geq p+2$, where T denotes the number of observations post initial condition, many transition probabilities in periods $t \in \{p+1, \dots, T-1\}$ admit at least two distinct transition functions. The combination of these two features motivates a two-step approach to obtain moment restrictions in panels of adequate length. In the first step, we shall compute the model transition functions. Then, the second step will simply consist in differencing two transition functions associated to the same transition probability. We show that a careful application of this procedure delivers all the moment equality restrictions available in the binary response case. We shall further elaborate on these steps in examples and use the resulting moment functions to derive new identification results. At a high level, the approach we advocate in this paper consists in solving a sequence of problems with identical structure period by period instead of solving directly a large system of equations based on the model full likelihood as in Honoré and Weidner (2020) and Dobronyi et al. (2021). As a consequence, our procedure remains tractable when the number of time periods increases and in models with higher order lags.

Besides the aforementioned papers, our work also connects to a line of research studying the identification of features of the distribution of fixed effects in discrete choice models. One branch in this literature has focused on developing general optimization tools to compute sharp numerical bounds on average marginal effects. This includes most notably the linear programming method of Honoré and Tamer (2006), recently adapted by Bonhomme et al. (2023) to the case of sequentially exogenous covariates, and the quadratic programming method of Chernozhukov et al. (2013). A second branch in this literature has sought instead to harness the specificities of logit models to obtain simple analytical bounds. In static logit models, Davezies et al. (2021) exploit mathematical results on the *moment problem* to formulate sharp bounds on the average partial effects of regressors on outcomes. In DFEL models, Aguirregabiria and Carro (2021) are the first to prove the point identification of average marginal effects in the baseline AR(1) logit model when $T \geq 3$. In related work, Dobronyi et al. (2021) make use of their moment equality and moment inequality restrictions to establish sharp bounds on functionals of the fixed effects such as average marginal effects and average posterior means in AR(1) and AR(2) specifications. We complement these results as a byproduct of our methodology: average marginal effects and their variants in AR(p) models, with arbitrary $p \geq 1$ are merely differences of average transition functions.

The remainder of the paper is organized as follows. Section 2 presents the setting and our main objective. Section 3 introduces some terminology and gives an outline of our procedure to construct moment restrictions. Section 4 implements our approach in AR(p) logit models with $p \geq 1$ and discusses identification of model parameters and average marginal effects. The semiparametric efficiency bound for the AR(1) is also characterized for an arbitrary number of time periods. Section 5 discusses extensions to the VAR(1) and the dynamic multinomial logit model with one lag, MAR(1) for short. In Section 6, we present an empirical illustration on the dynamics of drug consumption amongst young people and Section 7 offers concluding remarks. A complementary set of Monte Carlo simulations showing the small sample performance of GMM estimators based on our moment restrictions is available in Appendix Section D. Proofs are gathered in the Appendix.

2 Setting, assumptions and objective

Let $i = 1, \dots, N$ denote a population index and $t = 0, \dots, T$ be an index for time. We study DFEL models which may be viewed as threshold-crossing econometric specifications describing a discrete outcome Y_{it} through a latent index involving lagged outcomes (e.g Y_{it-1}), strictly exogenous regressors X_{it}, an individual-specific time-invariant unobservable A_i and an error term ϵ_{it}. The canonical example is the AR(1) model:

$$Y_{it} = \mathbb{1}\{\gamma_0 Y_{it-1} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\}, \quad t = 1, \dots, T$$

and we shall concentrate more broadly on cases where A_i is additively separable from the other explanatory variables. An initial condition that we will generically denote Y_i^0 completes such models to enable dynamics. The common parameter θ_0 is one target of interest and governs the influence of lagged outcomes and the regressors on the contemporaneous outcome. Other quantities of interest include counterfactual parameters such as average marginal effects.

Throughout, we leave the joint distribution of (Y_i^0, X_i, A_i) unrestricted where $X_i = (X_{i1}, \ldots, X_{iT})$ and thus refer to A_i as a fixed effect in common with the literature. The schocks ϵ_{it} are assumed to be serially independent logistically distributed, independent of (Y_i^0, X_i, A_i), except for the MAR(1) model where they are instead extreme value distributed. Finally, we shall assume that (Y_i, Y_i^0, X_i, A_i) are jointly i.i.d across individuals.

The data available to the econometrician consists of the initial condition Y_i^0, the outcome vector $Y_i = (Y_{i1}, \ldots, Y_{iT})$, and the covariates X_i for all N individuals. Interest centers primarily on the identification and estimation of θ_0 in short panels, i.e for fixed T. To this end, the chief objective of this paper is to show how to construct moment functions $\psi_\theta(Y_i, Y_i^0, X_i)$ free of the fixed effect parameter that are valid in the sense that:

$$\mathbb{E}\left[\psi_{\theta_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i, A_i\right] = 0 \tag{1}$$

When this is possible, the law of iterated expectations implies the conditional moment:

$$\mathbb{E}\left[\psi_{\theta_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i\right] = 0$$

which can in turn be leveraged to assess the identifiability of θ_0 and form the basis of a GMM estimation strategy. This is the central idea underlying functional differencing (Bonhomme (2012)) and was applied by Honoré and Weidner (2020) to derive valid moment conditions for a class of dynamic logit models with scalar fixed effects. We borrow the same insight but instead of searching for solutions numerically on a case-by-case basis, we propose a complementary systematic algebraic procedure to recover the model's valid moments [2]. In doing so, we flesh out the mechanics implied by the logistic assumption which in turn suggest a blueprint to deal with estimation of general DFEL models. For example, we are able

[2]Dobronyi et al. (2021) and Kitazawa (2022) also have an algebraic approach but our methodologies are very different. The first paper uses the full likelihood of the model and focuses on the AR(1) and special instances of the AR(2) model. The second paper has a transformation approach adapted to the AR(1) model. Our emphasis here is primarily on developing an approach that is tractable for a large class of models.

to characterize the expressions of valid moment functions in AR(p) models for arbitrary $p > 1$ which to the best of our knowledge is a new result in the literature. Furthermore, our approach carries over to multidimensional fixed effect specifications: VAR(1), dynamic network formation models and the MAR(1) in which searching for moments numerically is cumbersome or intractable.

In what follows, we shall use the shorthand $Y_{it_1}^{t_2} = (Y_{it_1}, \dots, Y_{it_2})$ to denote a collection of random variables over periods t_1 to t_2 with the convention that $Y_{it_1}^{t_2} = \emptyset$ if $t_1 > t_2$. Likewise, we may use the notation $y_{t_1}^{t_2} = (y_{t_1}, \dots, y_{t_2})$ to denote any $(t_2 - t_1)$-dimensional vector of reals with the convention $y_{t_1}^{t_2} = \emptyset$ for $t_1 > t_2$. Elements 1_n and 0_n shall refer to the n-dimensional vectors of ones and zeros respectively. The support of the outcome variable Y_{it} shall be denoted $\mathcal{Y}$. We let Δ denote the first-differencing operator so that $\Delta Z_{it} = Z_{it} - Z_{it-1}$ for any random variable Z_{it} and make use of the notation $Z_{its} = Z_{it} - Z_{is}$ for $s \neq t$ to accommodate long differences. We use $\mathbb{1}\{.\}$ for the indicator function; $\mathrm{Im}(f)$, $\ker(f)$, $\mathrm{rank}(f)$ to denote the image, the nullspace and the rank of a linear map f.

3 Outline of the procedure to derive valid moment functions

Let $T \geq 1$. Given an initial condition $y^0 \in \mathcal{Y}^p$, $p \geq 1$ being the lag order of the model, and strictly exogenous regressors $X_i \in \mathbb{R}^{K_x \times T}$, we denote the (one-period ahead) transition probability in period $t \geq 1$ from state $(l_1^t, y^0) \in \mathcal{Y}^t \times \mathcal{Y}^p$ to state $k \in \mathcal{Y}$ as:

$$\pi_t^{k|l_1^t, y^0}(A_i, X_i) = \pi_t^{k|l_1^t, y^0}(A_i, X_i; \theta_0) \equiv P(Y_{it+1} = k \mid Y_i^0 = y^0, Y_{i1}^t = l_1^t, X_i, A_i)$$

With p lags, the markovian nature of the models considered in this paper imply that $\pi_t^{k|l_1^t, y^0}(A_i, X_i)$ will not depend on the entire path of past outcomes but only on the value of the most recent p outcomes. For instance, in an AR(1) model where $p = 1$, we have:

$$\pi_t^{k|l_1^t, y^0}(A_i, X_i) = P(Y_{it+1} = k \mid Y_i^0 = y^0, Y_{i1}^t = l_1^t, X_i, A_i) = P(Y_{it+1} = k \mid Y_{it} = l_t, X_i, A_i)$$

and thus we will suppress the dependence on $(y^0, l_1, \ldots, l_{t-1})$ and write $\pi_t^{k|l_t}(A_i, X_i)$. We shall proceed analogously for the more general case $p \geq 1$.

We call a *transition function* associated to a transition probability $\pi_t^{k|l_1^t, y^0}(A_i, X_i)$ any moment function $\phi_\theta^{k|l_1^t, y^0}(Y_i, Y_i^0, X_i)$ of the data and the common parameters verifying:

$$\mathbb{E}\left[\phi_{\theta_0}^{k|l_1^t, y_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i, A_i\right] = \pi_t^{k|l_1^t, y_0}(A_i, X_i) \tag{2}$$

With these notions in hand, we are ready to describe our two-step approach to derive valid moment functions in the sense of equation (1). In **Step 1)**, we begin by computing the model's transition functions. Our procedure requires a minimum of $T = p + 1$ periods of observations to accommodate arbitrary regressors and initial condition. In this case, we can get analytical formulas for the transition functions associated to the transition probabilities in period $t = p$ and Theorem 1 and Theorem 3 below imply that they are unique. However, this is not immediately helpful to get moment (equality) restrictions on θ_0. We require one more period. As soon as $T \geq p+2$, we explain how to construct distinct transition functions associated to the same transition probabilities in periods $t \in \{p+1, \ldots, T-1\}$. The key ingredient is the use of *partial fraction decompositions* for *rational fractions* adapted to the structure of the transition probabilities. It is then a matter of taking differences of two transition functions associated to the same transition probability to obtain valid moment functions; we refer to this last step as **Step 2)**. The ensuing sections demonstrate this procedure in scalar and multidimensional fixed effect models.

4 Scalar fixed effect models

4.1 Moment restrictions for the AR(1) logit model

For exposition, we begin with the baseline AR(1) logit model with fixed effects introduced above:

$$Y_{it} = \mathbb{1}\{\gamma_0 Y_{it-1} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\}, \quad t = 1, \ldots, T \tag{3}$$

Here, $\mathcal{Y} = \{0,1\}$, $\theta_0 = (\gamma_0, \beta_0') \in \mathbb{R} \times \mathbb{R}^{K_x}$, the initial condition Y_i^0 consists of the binary-valued random variable Y_{i0} and $A_i \in \mathbb{R}$.

4.1.1 The number of moment restrictions in the AR(1)

We start out by enumerating the moment restrictions implied by the model. This will provide a means to assess the exhaustiveness of our approach. To this end, let $\mathcal{E}_{y_0,x}$ denote the conditional expectation operator mapping any function of the outcome variable Y_i to its conditional expectation given $Y_{i0} = y_0, X_i = x$ and the fixed effect A_i, i.e

$$\begin{aligned} \mathcal{E}_{y_0,x} \colon \mathbb{R}^{\mathcal{Y}^T} &\longrightarrow \mathbb{R}^{\mathbb{R}} \\ \phi(.;y_0,x) &\longmapsto \mathbb{E}\left[\phi(Y_i,y_0,x)|Y_{i0} = y_0, X_i = x, A_i = .\right] \end{aligned}$$

For example, for any $y \in \mathcal{Y}^T$, $\mathcal{E}_{y_0,x}\left[\mathbb{1}\{. = y\}\right]$ yields the conditional probability of observing history y for all possible values of the fixed effect, i.e:

$$\mathcal{E}_{y_0,x}\left[\mathbb{1}\{. = y\}\right] = P(Y_i = y|Y_{i0} = y_0, X_i = x, A_i = .)$$

where $P(Y_i = y|Y_{i0} = y_0, X_i = x, A_i = a) = \prod_{t=1}^{T} \frac{e^{y_t(\gamma_0 y_{t-1} + x_t'\beta_0 + a)}}{1+e^{\gamma_0 y_{t-1} + x_t'\beta_0 + a}}, \quad \forall a \in \mathbb{R}$. Then, we have the following result,

Theorem 1. *Consider model (3) with $T \geq 1$ and initial condition $y_0 \in \mathcal{Y}$. Suppose that for any $t, s \in \{1, \ldots, T-1\}$ and $y, \tilde{y} \in \mathcal{Y}$, $\gamma_0 y + x_t'\beta_0 \neq \gamma_0 \tilde{y} + x_s'\beta_0$ if $t \neq s$ or $y \neq \tilde{y}$. Then, the family $\mathcal{F}_{y_0,T} = \left\{1, \pi_0^{y_0|y_0}(.,x), (\pi_t^{0|0}(.,x), \pi_t^{1|1}(.,x))_{t=1}^{T-1}\right\}$ of size $2T$ forms a basis of $\mathrm{Im}(\mathcal{E}_{y_0,x})$ and $\dim\left(\ker(\mathcal{E}_{y_0,x})\right) = 2^T - 2T$.*

Theorem 1 formalizes the intuition that the transition probabilities summarize the parametric component of the model: 2^T histories are possible yet only $2T$ basis elements are necessary to fully characterize their conditional probabilities. This follows from the observation that when the covariate index [3] of each transition probability differ, the conditional probability of each history $y \in \mathcal{Y}^T$ is a ratio of polynomials in e^a, where the numerator has lower degree than the denominator, and the later is a product of distinct irreducible terms. A sufficient condition for this is that $\gamma_0 \neq 0$ and that one regressor is continuously distributed with

[3] We refer to the quantity $\gamma_0 y_{t-1} + x_t'\beta_0$ for a given period t.

non-zero slope. In turn, standard results on *partial fraction decompositions* ensure that this ratio can be expressed as a unique linear combination of transition probabilities. To finally conclude that $\mathcal{F}_{y_0,T}$ is a basis of $\text{Im}(\mathcal{E}_{y_0,x})$, we leverage upcoming results demonstrating that the transition probabilities live in $\text{Im}(\mathcal{E}_{y_0,x})$ as expectations of transition functions.

Importantly, since $\ker(\mathcal{E}_{y_0,x})$ is the set of valid moment functions verifying equation (1), Theorem 1 tells us that the AR(1) model features $2^T - 2T$ linearly independent moment restrictions in general. This is a consequence of the *rank nullity theorem* for linear maps with finite dimensional domains. The fact that $2^T - 2T$ moment conditions are available for the AR(1) appeared initially as a conjecture in Honoré and Weidner (2020) and was later established by Kruiniger (2020) and Dobronyi et al. (2021) using different arguments from here. They do not emphasize the role of the transition probabilities. Our ideas extend naturally to the case of arbitrary lags which was hitherto an open problem. We discuss this extension in Section 4.4.1.

Remark 1 (Counting moments in logit models)**.** The idea of decomposing the conditional probabilities of all choice histories in a basis provides a useful device to infer a lower bound on the number of moment restrictions in logit models. If one can further prove that elements of this basis belong to the image of the conditional expectation operator, then this lower bound coincides with the exact number of moment restrictions.

- In the static panel logit model of Rasch (1960), $\gamma_0 = 0$ and we have $\pi_t^{1|1}(.,x) = 1 - \pi_t^{0|0}(.,x)$. Thus, provided that $x_t'\beta_0 \neq x_s'\beta_0$ for all $t \neq s$, $\mathcal{F}_T = \left\{1, (\pi_t^{0|0}(.,x))_{t=0}^{T-1}\right\}$ spans the image of the conditional expectation operator. This implies at least $2^T - (T+1)$ moment restrictions. It turns out that $2^T - (T+1)$ is precisely the total number of moment restrictions for this model. This follows from Remark 6 below which characterizes the transition functions associated to each element of $\mathcal{F}_T$.

- In the Cox (1958) model, $\gamma_0 \neq 0$ and $\beta_0 = 0$ and the transition probabilities are: $\pi^{0|0}(a) = \frac{1}{1+e^a}$ and $\pi^{1|1}(a) = \frac{e^{\gamma_0+a}}{1+e^{\gamma_0+a}}$ (or equivalently $\pi^{0|1}(a) = \frac{1}{1+e^{\gamma_0+a}}$). See the next section for further details. In this case, the family $\mathcal{F}_{y_0,T} = \left\{1, \left(\pi^{0|0}(.)^j, \pi^{0|1}(.)^j\right)_{j=1}^{T-1}, \pi^{0|y_0}(.)^T\right\}$ which consists of powers of the time-invariant transition probabilities spans the image of the conditional expectation operator. Since

$|\mathcal{F}_{y_0,T}| = 2T$, the model produces at least $2^T - 2T$ linearly independent moment restrictions.

Remark 2 (A matrix perspective)**.** Since $\mathcal{E}_{y_0,x}$ is a linear map, it admits a unique $2^T \times 2T$ matrix representation $\Lambda_{y_0,x}$ where each row translates the conditional probability of a choice history $y \in \mathcal{Y}^T$ in terms of the transition probabilities of $\mathcal{F}_{y_0,T}$[4]. From this point of view, valid moments correspond to 2^T-vectors ψ in the left nullspace of $\Lambda_{y_0,x}$, meaning $\psi'\Lambda_{y_0,x} = 0$. Constructing $\Lambda_{y_0,x}$ and then solving this $2T$ linear system of equations in 2^T unknowns directly is straightforward using symbolic tools when T is "small" (e.g Dobronyi et al. (2021), Honoré and Weidner (2020)) but is computationally impractical otherwise. Instead, we propose a constructive approach to back out analytic expressions of the valid moment functions that is tractable for arbitrary values of T.

Having clarified the total count of moment restrictions in the AR(1) logit model, we next discuss how to construct them with our two-step procedure.

4.1.2 Construction of valid moment functions for the pure model

In the absence of exogenous regressors, model (3) simplifies to:

$$Y_{it} = \mathbb{1}\{\gamma_0 Y_{it-1} + A_i - \epsilon_{it} \geq 0\}, \quad t = 1, \ldots, T \tag{4}$$

which was first introduced by Cox (1958) and then revisited in Chamberlain (1985), Magnac (2000). These papers established the identification of γ_0 for $T \geq 3$ via conditional likelihood based on the insight that $(Y_{i0}, \sum_{t=1}^{T-1} Y_{it}, Y_{iT})$ are sufficient statistics for the fixed effect. Our methodology is conceptually different as we seek to directly construct moment functions verifying equation (1).

For what follows, it is helpful to remember that the individual-specific transition probability from state l to state k is time-invariant and given by:

$$\pi^{k|l}(A_i) = P(Y_{it+1} = k | Y_{it} = l, A_i) = \frac{e^{k(\gamma_0 l + A_i)}}{1 + e^{\gamma_0 l + A_i}}, \quad \forall (l,k) \in \mathcal{Y}$$

[4]Entries of this matrix may be found using for example the identities in Appendix Lemma 8 or any other standard textbook tools for *rational fractions*.

Step 1). We shall begin by deriving the transition functions for $\pi^{0|0}(A_i)$ and $\pi^{1|1}(A_i)$. Observe that $\pi^{1|0}(A_i)$ and $\pi^{0|1}(A_i)$ are effectively redundant since probabilities sum to one. A natural starting place is to investigate the case $T = 2$, i.e 2 periods of observations after the initial condition. Recalling definition (2), we search for $\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0})$, respectively $\phi_\theta^{1|1}(Y_{i2}, Y_{i1}, Y_{i0})$, whose conditional expectation given (Y_{i0}, A_i) yields $\pi^{0|0}(A_i)$, respectively $\pi^{1|1}(A_i)$. For the purposes of illustration and to show the kind of calculations arising broadly in DFEL models, let us derive $\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0})$. By Bayes's rule:

$$
\begin{aligned}
&\mathbb{E}\left[\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0}) \mid Y_{i0} = y_0, A_i = a\right] \\
&\quad = \sum_{y_2=0}^{1}\sum_{y_1=0}^{1} P(Y_{i2} = y_2|Y_{i1} = y_1, A_i = a)P(Y_{i1} = y_1|Y_{i0} = y_0, A_i = a)\phi_\theta^{0|0}(y_2, y_1, y_0) \\
&\qquad\quad = \frac{e^{\gamma_0 y_0 + a}}{1 + e^{\gamma_0 y_0 + a}}\left(\frac{e^{\gamma_0 + a}}{1 + e^{\gamma_0 + a}}\phi_\theta^{0|0}(1, 1, y_0) + \frac{1}{1 + e^{\gamma_0 + a}}\phi_\theta^{0|0}(0, 1, y_0)\right) \\
&\qquad\qquad + \frac{1}{1 + e^{\gamma_0 y_0 + a}}\left(\frac{e^{a}}{1 + e^{a}}\phi_\theta^{0|0}(1, 0, y_0) + \frac{1}{1 + e^{a}}\phi_\theta^{0|0}(0, 0, y_0)\right)
\end{aligned}
$$

where the second equality uses the logistic hypothesis. By quick inspection, we see that the terms in the first parenthesis have $(1+e^{\gamma_0+a})$ in their denominator unlike $\pi^{0|0}(A_i)$. Because $-e^{-\gamma_0}$ is not a *pole* of $\pi^{0|0}(A_i)$[5], we conclude that $\phi_\theta^{0|0}(1, 1, y_0) = \phi_\theta^{0|0}(0, 1, y_0) = 0$. This first deduction leaves us with

$$
\mathbb{E}\left[\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0}) \mid Y_{i0} = y_0, A_i = a\right] = \frac{1}{1 + e^{\gamma_0 y_0 + a}}\left(\frac{e^{a}}{1 + e^{a}}\phi_\theta^{0|0}(1, 0, y_0) + \frac{1}{1 + e^{a}}\phi_\theta^{0|0}(0, 0, y_0)\right)
$$

Now, since $\pi^{0|0}(A_i)$ does not depend on y_0, we must cancel the denominator $(1 + e^{\gamma_0 y_0 + a})$. To achieve this, we must set: $\phi_{\theta_0}^{0|0}(1, 0, y_0) = C_0 e^{\gamma_0 y_0}, \phi_{\theta_0}^{0|0}(0, 0, y_0) = C_0$ for some constant $C_0 \in \mathbb{R} \setminus \{0\}$. Then,

$$
\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{i2}, Y_{i1}, Y_{i0})|Y_{i0} = y_0, A_i = a\right] = C_0\frac{1}{1 + e^{a}}
$$

and $C_0 = 1$ is the appropriate normalization to obtain the desired transition function. Of course, the exact same logic applies for $\phi_{\theta_0}^{1|1}(Y_{i2}, Y_{i1}, Y_{i0})$ and $\pi^{1|1}(A_i)$.

This short calculation provides a useful recipe for the general case $T \geq 2$. We learned

[5] A *pole* of a rational function is a root of its denominator. Formally, we are substituting $u = e^a$ and we are extending $\pi^{0|0}(u)$ to the real line.

that we can search for functions of three consecutive outcomes $\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1})$ such that:

$$\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}) = \mathbb{1}\{Y_{it} = k\}\phi_\theta^{k|k}(Y_{it+1}, k, Y_{it-1})$$
$$\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}) \mid Y_{i0}, Y_{i1}^{t-1}, A_i\right] = \pi^{k|k}(A_i)$$

The first restriction is a functional form that eliminates terms with inadequate *poles* after taking expectations. The second restriction is a normalization condition to match the desired transition probability. Following this argument, we arrive at the expressions in Lemma 1.

Lemma 1. *In model (4) with $T \geq 2$ and $t \in \{1, \dots, T-1\}$, let*

$$\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}) = (1 - Y_{it})e^{\gamma Y_{it+1} Y_{it-1}}$$
$$\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}) = Y_{it}e^{\gamma(1-Y_{it+1})(1-Y_{it-1})}$$

Then:

$$\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1})|Y_{i0}, Y_{i1}^{t-1}, A_i\right] = \pi^{0|0}(A_i) = \frac{1}{1 + e^{A_i}}$$
$$\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1})|Y_{i0}, Y_{i1}^{t-1}, A_i\right] = \pi^{1|1}(A_i) = \frac{e^{\gamma_0 + A_i}}{1 + e^{\gamma_0 + A_i}}$$

Remark 3 (Connection to Kitazawa)**.** Interestingly, Lemma 1 is a reformulation of results first shown by Kitazawa et al. (2013, 2016), Kitazawa (2022), albeit with a very different logic than the calculations displayed above. We set out the connection between our respective approaches in Section 4.3 where we also discuss the case with exogenous regressors.

Step 2). The second step in the agenda is the construction of valid moment functions. Because the transition probability of the model are time-invariant, one trivial way to achieve this is to consider the pairwise difference of $\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1})$ and $\phi_\theta^{k|k}(Y_{is+1}, Y_{is}, Y_{is-1})$ for any feasible $s \neq t$. This is the content of Proposition 1. We will need a minimum of four total periods of observations, which coincides with the requirements of the conditional likelihood approach.

Proposition 1. *In model (4) with $T \geq 3$, let*

$$\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s+1}) = \phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}) - \phi_\theta^{k|k}(Y_{is+1}, Y_{is}, Y_{is-1})$$

for all $k \in \mathcal{Y}$*,* $t \in \{2, \dots, T-1\}$ *and* $s \in \{1, \dots, t-1\}$*. Then,*

$$\mathbb{E}\left[\psi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s+1})|Y_{i0}, Y_{i1}^{s-1}, A_i\right] = 0$$

Remark 4 (Efficient GMM)**.** Given that the conditional likelihood is semi-parametrically efficient for $T = 3$ (Gu et al. (2023), Hahn (2001)), it is natural to ask whether the approach advocated here accounts for all the information in the model in that case. It turns out that it does. Specifically, letting $s_i^c(\theta)$ denote the conditional scores when $y_0 = 0$ as in Hahn (2001), we have:

$$s_i^c(\gamma_0) = \frac{1}{(1+e^{\gamma_0})(e^{-\gamma_0}-1)}\left(\psi_\theta^{0|0}(Y_{i1}^3, Y_{i1}^2, 0) + \psi_\theta^{1|1}(Y_{i1}^3, Y_{i1}^2, 0)\right)$$

where the right-hand side corresponds to the efficient moment for the moment restriction $\mathbb{E}\left[\psi_\theta(Y_{i1}^3, Y_{i0}^2)|Y_{i0} = 0\right] = 0$, $\psi_\theta(Y_{i1}^3, Y_{i1}^2, 0) = (\psi_\theta^{0|0}(Y_{i1}^3, Y_{i1}^2, 0), \psi_\theta^{1|1}(Y_{i1}^3, Y_{i1}^2, 0))'$.

4.1.3 Construction of valid moment functions with strictly exogenous regressors

In this subsection, we move on to the AR(1) logit model with strictly exogenous covariates characterized by equation (3).

Step 1). We employ the same shortcut recipe as in the "pure" case and begin by looking for moment functions $\phi_\theta^{0|0}(.)$ and $\phi_\theta^{1|1}(.)$ verifying:

$$\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = \mathbb{1}\{Y_{it} = k\}\phi_\theta^{k|k}(Y_{it+1}, k, Y_{it-1}, X_i)$$
$$\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i), \quad k \in \mathcal{Y}$$

where this time

$$\pi_t^{k|l}(A_i, X_i) = P(Y_{it+1} = k|Y_{it} = l, X_i, A_i) = \frac{e^{k(\gamma_0 l + X_{it+1}'\beta_0 + A_i)}}{1 + e^{\gamma_0 l + X_{it+1}'\beta_0 + A_i}}, \quad \forall (k, l) \in \mathcal{Y}^2$$

The same simple calculations described just above lead to the expressions in Lemma 2. The only (expected) change is the appearance of a new term $+/-\Delta X_{it+1}'\beta$ which accounts for the presence of covariates in the model.

Lemma 2. *In model (3) with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, let*

$$\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = (1 - Y_{it})e^{Y_{it+1}\left(\gamma Y_{it-1} - \Delta X'_{it+1}\beta\right)}$$
$$\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = Y_{it}e^{(1-Y_{it+1})\left(\gamma(1-Y_{it-1}) + \Delta X'_{it+1}\beta\right)}$$

Then:

$$\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{0|0}(A_i, X_i) = \frac{1}{1 + e^{A_i + X'_{it+1}\beta_0}}$$
$$\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{1|1}(A_i, X_i) = \frac{e^{\gamma_0 + X'_{it+1}\beta_0 + A_i}}{1 + e^{\gamma_0 + X'_{it+1}\beta_0 + A_i}}$$

At this point, it is important to highlight that unlike previously, the transition probabilities are covariate-dependent. The upshot is that the naive difference of $\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)$ and $\phi_\theta^{k|k}(Y_{is+1}, Y_{is}, Y_{is-1}, X_i)$ for $s \neq t$ no longer leads to valid moment functions in general. Indeed, while Lemma 2 ensures that

$$\mathbb{E}\left[\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) - \phi_\theta^{k|k}(Y_{is+1}, Y_{is}, Y_{is-1}, X_i)|Y_{i0}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i) - \pi_s^{k|k}(A_i, X_i)$$

clearly, $\pi_t^{k|k}(A_i, X_i) - \pi_s^{k|k}(A_i, X_i) \neq 0$ when $X'_{it+1}\beta_0 \neq X'_{is+1}\beta_0$ [6]. Thus, a different logic is required in the presence of explanatory variables other than a first order lag.

The key, as foreshadowed in Section 3 is that as soon as $T \geq 3$, it is possible to construct transition functions other than $\phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i)$ also mapping to $\pi_t^{k|k}(A_i, X_i)$ in time periods $t \in \{2, \ldots, T-1\}$. These new transition functions that we denote $\zeta_\theta^{k|k}(.)$ to emphasize their difference have a particular form. They consist of a weighted combination of past outcome $\mathbb{1}(Y_{is} = k)$, $1 \leq s < t$, and the interaction of $\mathbb{1}(Y_{is} \neq k)$ with any transition function associated to $\pi_t^{k|k}(A_i, X_i)$ having no dependence on outcomes prior to period s, e.g $\phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i)$. This property follows from a *partial fraction decomposition* presented in Lemma 8 that exploits the structure of the model probabilities under the logistic assumption. It relates to the hyperbolic transformations ideas of Kitazawa (2022). In the sequel, we shall see that this insight carries over to the AR(p) logit model with $p > 1$. Lemma 3 below gives the "simplest" additional transition functions that one can construct when $T \geq 3$ for the

[6] A matching strategy in the spirit of Honoré and Kyriazidou (2000) may still be applicable when in our example $X_{it+1} = X_{is+1}$. However, this is known to lead to estimators converging at rate less than $\sqrt{N}$ for continuous covariates and it rules out certain regressors such as time dummies and time trends.

AR(1) model with exogenous regressors (the only ones when $T = 3$).

Lemma 3. *In model (3) with $T \geq 3$, for all t, s such that $T-1 \geq t > s \geq 1$, let:*

$$
\begin{aligned}
\mu_s(\theta) &= \gamma Y_{is-1} + X_{is}'\beta \\
\kappa_t^{0|0}(\theta) &= X_{it+1}'\beta, \quad \kappa_t^{1|1}(\theta) = \gamma + X_{it+1}'\beta \\
\omega_{t,s}^{0|0}(\theta) &= 1 - e^{(\kappa_t^{0|0}(\theta) - \mu_s(\theta))}, \quad \omega_{t,s}^{1|1}(\theta) = 1 - e^{-(\kappa_t^{1|1}(\theta) - \mu_s(\theta))}
\end{aligned}
$$

and define the moment functions:

$$
\begin{aligned}
\zeta_\theta^{0|0}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) &= (1 - Y_{is}) + \omega_{t,s}^{0|0}(\theta) Y_{is} \phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) \\
\zeta_\theta^{1|1}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) &= Y_{is} + \omega_{t,s}^{1|1}(\theta)(1 - Y_{is}) \phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)
\end{aligned}
$$

Then,

$$
\begin{aligned}
\mathbb{E}\left[\zeta_{\theta_0}^{0|0}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) | Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] &= \pi_t^{0|0}(A_i, X_i) \\
\mathbb{E}\left[\zeta_{\theta_0}^{1|1}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) | Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] &= \pi_t^{1|1}(A_i, X_i)
\end{aligned}
$$

When $T \geq 4$, it turns out that we can build even more transition functions from those given in Lemma 3 by repeating the same type of logic based on *partial fraction expansions*; Corollary 3.1 provides a recursive formulation.

Corollary 3.1. *In model (3) with $T \geq 4$, for any t and ordered collection of indices s_1^J, $J \geq 2$, satisfying $T - 1 \geq t > s_1 > \ldots > s_J \geq 1$, let*

$$
\begin{aligned}
\zeta_\theta^{0|0}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) &= (1 - Y_{is_J}) + \omega_{t,s_J}^{0|0}(\theta) Y_{is_J} \zeta_\theta^{0|0}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_{J-1}-1}^{s_{J-1}}, X_i) \\
\zeta_\theta^{1|1}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) &= Y_{is_J} + \omega_{t,s_J}^{1|1}(\theta)(1 - Y_{is_J}) \zeta_\theta^{1|1}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_{J-1}-1}^{s_{J-1}}, X_i)
\end{aligned}
$$

with weights $\omega_{t,s_J}^{0|0}(\theta), \omega_{t,s_J}^{1|1}(\theta)$ defined as in Lemma 3. Then,

$$
\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) | Y_{i0}, Y_{i1}^{s_J-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i), \quad \forall k \in \mathcal{Y}
$$

Step 2). Provided $T \geq 3$, the difference between any transition functions associated to the same transition probabilities in periods $t \in \{2, \ldots, T-1\}$ constitutes a valid candidate for (1). One particularly relevant set of valid moment functions for reasons explained below is presented in Proposition 2.

Proposition 2. *In model (3), for all $k \in \mathcal{Y}$,*
if $T \geq 3$, for all t, s such that $T-1 \geq t > s \geq 1$, let

$$\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) = \phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i) - \zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i),$$

if $T \geq 4$, for any t and ordered collection of indices s_1^J, $J \geq 2$, satisfying $T-1 \geq t > s_1 > \ldots > s_J \geq 1$, let

$$\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) = \phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i) - \zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i),$$

Then,

$$\mathbb{E}\left[\psi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] = 0$$
$$\mathbb{E}\left[\psi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i)|Y_{i0}, Y_{i1}^{s_J-1}, X_i, A_i\right] = 0$$

This family of moment functions has cardinality $2^T - 2T$ which by Theorem 1 is precisely the number of linearly independent moment conditions available for the AR(1). To see this, notice that for fixed $(k, Y_{i0}) \in \mathcal{Y}^2$, and a given time period $t \in \{2, \ldots, T-1\}$, Proposition 2 gives a total of:

$$\sum_{l=1}^{t-1} \binom{t-1}{l} = 2^{t-1} - 1$$

valid moment functions. This follows from a simple counting argument. First, we get $\binom{t-1}{1}$ possibilities from choosing any s in $\{1, \ldots, t-1\}$ to form $\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)$. To that, we must add another $\sum_{l=2}^{t-1}\binom{t-1}{l}$ possibilities from choosing all feasible sequences s_1^J with $t-1 \geq s_1 > s_2 > \ldots > s_J \geq 1$ to form $\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i)$. Summing over $t = 2, \ldots, T-1$ and multiplying by 2 to account for the two possible values for k delivers the result:

$$2 \times \sum_{t=2}^{T-1}\sum_{l=1}^{t-1} \binom{t-1}{l} = 2 \times \sum_{t=2}^{T-1}(2^{t-1} - 1) = 2^T - 2T$$

Furthermore, there is evidence that the family is linearly independent. It is readily verified for

$T = 3$ since the two valid moment functions produced by the model depend on two distinct sets of choice histories. This can be seen from their unpacked expressions in equations (9) and (10) in the Appendix. Unfortunately, this argument does not carry over to longer panels but we have verified numerically that the linear independence property of this family continues to hold for several different values of $T \geq 4$. This suggests that our approach delivers all the *moment equality* restrictions available in the AR(1) model with T periods post initial condition [7].

Remark 5 (Symmetry)**.** The transition functions and valid moment functions of the AR(1) model share a special symmetry property. Indeed, by inspection the transition functions of Lemma 2 verify

$$\phi_\theta^{0|0}(1 - Y_{it+1}, 1 - Y_{it}, 1 - Y_{it-1}, -X_i) = \phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)$$

It is not difficult to see that this symmetry, i.e substituting Y_{it} by $(1 - Y_{it})$ and X_{it} by $-X_{it}$ to obtain $\phi_\theta^{1|1}(Y_{it-1}^{t+1}, X_i)$ from $\phi_\theta^{0|0}(Y_{it-1}^{t+1}, X_i)$ transfers to the other transition functions of Lemma 3, Corollary 3.1 and ultimately to the valid moment functions of Proposition 2. This symmetry can be useful for computational purposes.

Remark 6 (Static logit)**.** If $\gamma_0 = 0$, model (3) specializes to the static panel logit model of Rasch (1960) and our two-step approach is still applicable. For that case, Lemma 2 gives two moment functions for $T = 2$:

$$\begin{aligned}
\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, X_i) &= (1 - Y_{i1})e^{-Y_{i2}\Delta X_2'\beta} \\
\phi_\theta^{1|1}(Y_{i2}, Y_{i1}, X_i) &= Y_{i1}e^{(1-Y_{i2})\Delta X_{i2}'\beta}
\end{aligned}$$

such that $\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{i1}^2, , X_i)|X_i, A_i\right] = \frac{1}{1+e^{X_{i2}'\beta_0+A_i}}$ and $\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{i1}^2, X_i)|X_i, A_i\right] = \frac{e^{X_{i2}'\beta_0+A_i}}{1+e^{X_{i2}'\beta_0+A_i}}$. It follows that a valid moment function with two periods of observation is

$$\begin{aligned}
\psi_\theta(Y_{i2}, Y_{i1}, X_i) &= \phi_\theta^{1|1}(Y_{i2}, Y_{i1}, X_i) - (1 - \phi_\theta^{0|0}(Y_{i2}, Y_{i1}, X_i)) \\
&= (1 - e^{-\Delta X_{i2}'\beta})\left(Y_{i1}(1 - Y_{i2})e^{\Delta X_{i2}'\beta} - (1 - Y_{i1})Y_{i2}\right)
\end{aligned}$$

[7] This is not all the identifying content of the AR(1) specification since we know from Dobronyi et al. (2021) that the model also implies moment inequality conditions.

which is proportional to the score of the conditional likelihood based on the sufficient statistic $Y_{i1} + Y_{i2}$ (Rasch (1960), Andersen (1970), Chamberlain (1980)).

4.2 Semiparametric efficiency bound for the AR(1) with regressors

Honoré and Weidner (2020) gave sufficient conditions to identify $\theta_0 = (\gamma_0, \beta_0')'$ in the AR(1) model with $T \geq 3$. A natural follow-up question is to ask how accurately can θ_0 be estimated in that case, or equivalently what is the semiparametric efficiency bound. In a corrigendum to Hahn (2001), Gu et al. (2023) confirmed that the conditional likelihood estimator is semiparametrically efficient for $T = 3$ in the "pure" AR(1) model. However, the characterization of the semiparametric efficiency bound and the question of what estimator attains it generally remains unclear with covariates and potentially more time periods.

To answer these questions, let $\psi_\theta(Y_{i0}, Y_i, X_i)$ denote the $(2^T - 2T)$-vector stacking all the valid moment functions of Proposition 2 for an arbitrary number of time periods $T \geq 3$. Additionally, let $D(Y_{i0}, X_i) = \mathbb{E}\left[\frac{\partial \psi_{\theta_0}(Y_{i0}, Y_i, X_i)}{\partial \theta'} | Y_{i0}, X_i\right]$ and let $\Sigma(Y_{i0}, X_i) = \mathbb{E}\left[\psi_{\theta_0}(Y_{i0}, Y_i, X_i)\psi_{\theta_0}(Y_{i0}, Y_i, X_i)' | Y_{i0}, X_i\right]$.

Assumption 1. *In model (3) with $T \geq 3$, assume that i) $\mathbb{E}\left[X_i X_i'\right] < \infty$ and ii) the matrix $\mathbb{E}\left[D(Y_{i0}, X_i)'\Sigma(Y_{i0}, X_i)^{-1}D(Y_{i0}, X_i)\right]$ exists and is nonsingular.*

With these notations in hand and under the conditions of Assumption 1, Theorem 2 clarifies that the *efficient score* coincides with the efficient moment for the conditional moment problem: $\mathbb{E}\left[\psi_\theta(Y_{i0}, Y_i, X_i) | Y_{i0}, X_i\right] = 0$. Put differently, the maximal efficiency with which θ_0 can be estimated is $V_0 = \mathbb{E}[D(Y_{i0}, X_i)'\Sigma(Y_{i0}, X_i)^{-1}D(Y_{i0}, X_i)]^{-1}$. This result is in accordance with Remark 4 which noted that the score of the conditional likelihood without covariates is precisely the efficient moment implied by our conditional moment restrictions in this case.

Theorem 2. *Consider model (3) with $T \geq 3$ and suppose that Assumption 1 holds. Then, the semiparametric variance bound for θ_0 is finite and given by*
$V_0 = \mathbb{E}[D(Y_{i0}, X_i)'\Sigma(Y_{i0}, X_i)^{-1}D(Y_{i0}, X_i)]^{-1}$.

The proof of Theorem 2 consists in a verification of the conditions for an application of Theorem 3.2 in Newey (1990). Interestingly, Davezies et al. (2023) presented analogous results in the static panel data case with three periods of observations.

4.3 Connections to other works on the AR(1) logit model

As indicated previously, there is a connection between our methodology and that of Kitazawa (2022) for the AR(1) model. Indeed, after some algebraic manipulation, we can re-express the transition functions of Lemma 2 (or Lemma 1 without covariates) as:

$$\phi_\theta^{0|0}(Y_{it-1}^{t+1}, X_i) = 1 - Y_{it} - (1-Y_{it})Y_{it+1} + (1-Y_{it})Y_{it+1}e^{-\Delta X_{it+1}'\beta} + \delta Y_{it-1}(1-Y_{it+1})Y_{it+1}e^{-\Delta X_{it+1}'\beta}$$
$$\phi_\theta^{1|1}(Y_{it-1}^{t+1}, X_i) = Y_{it}Y_{it+1} + Y_{it}(1-Y_{it+1})e^{\Delta X_{it+1}'\beta} + \delta(1-Y_{it-1})Y_{it}(1-Y_{it+1})e^{\Delta X_{it+1}'\beta}$$

where $\delta = (e^\gamma - 1)$. Thus, the moment conditions of Lemma 2 imply that we can write:

$$Y_{it} + (1-Y_{it})Y_{it+1} - (1-Y_{it})Y_{it+1}e^{-\Delta X_{it+1}'\beta_0} - \delta_0 Y_{it-1}(1-Y_{it+1})Y_{it+1}e^{-\Delta X_{it+1}'\beta_0} = \frac{e^{X_{it+1}'\beta_0 + A_i}}{1 + e^{X_{it+1}'\beta_0 + A_i}} + \epsilon_{it}^{0|0}$$
$$Y_{it}Y_{it+1} + Y_{it}(1-Y_{it+1})e^{\Delta X_{it+1}'\beta_0} + \delta_0(1-Y_{it-1})Y_{it}(1-Y_{it+1})e^{\Delta X_{it+1}'\beta_0} = \frac{e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}}{1 + e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}} + \epsilon_{it}^{1|1}$$

where $\mathbb{E}\left[\epsilon_{it}^{0|0}|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = 0$ and $\mathbb{E}\left[\epsilon_{it}^{1|1}|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = 0$. These expressions are the so-called *h-form* and *g-form* of Kitazawa (2022) for model (3) and were originally obtained through an ingenious usage of the mathematical properties of the hyperbolic tangent function. The evident connection between the transition functions and the *h-form* and *g-form* offers an interesting new perspective on the transformation approach of Kitazawa (2022) for the AR(1) model. If we further define

$$U_{it} = Y_{it} + (1-Y_{it})Y_{it+1} - (1-Y_{it})Y_{it+1}e^{-\Delta X_{it+1}'\beta} - \delta Y_{it-1}(1-Y_{it+1})Y_{it+1})e^{-\Delta X_{it+1}'\beta}$$
$$\Upsilon_{it} = Y_{it}Y_{it+1} + Y_{it}(1-Y_{it+1})e^{\Delta X_{it+1}'\beta} + \delta(1-Y_{it-1})Y_{it}(1-Y_{it+1})e^{\Delta X_{it+1}'\beta}$$

the two moment functions of Kitazawa (2022) for the AR(1) model write

$$\hbar U_{it} = U_{it} - Y_{it-1} - \tanh\left(\frac{-\gamma Y_{it-2} + (\Delta X_{it} + \Delta X_{it+1})'\beta}{2}\right)(U_{it} + Y_{it-1} - 2U_{it}Y_{it-1})$$
$$\hbar \Upsilon_{it} = \Upsilon_{it} - Y_{it-1} - \tanh\left(\frac{\gamma(1-Y_{it-2}) + (\Delta X_{it} + \Delta X_{it+1})'\beta}{2}\right)(\Upsilon_{it} + Y_{it-1} - 2\Upsilon_{it}Y_{it-1})$$

which can be formulated in terms of our own moment functions as

$$
\begin{aligned}
\hbar U_{it} &= -\frac{2}{2-\omega_{t,t-1}^{0|0}(\theta)}\psi_{\theta}^{0|0}(Y_{it-1}^{t+1}, Y_{it-2}^{t-1}, X_i) \\
\hbar \Upsilon_{it} &= \frac{2}{2-\omega_{t,t-1}^{1|1}(\theta)}\psi_{\theta}^{1|1}(Y_{it-1}^{t+1}, Y_{it-2}^{t-1}, X_i)
\end{aligned}
$$

Appendix Section B provides detailed derivations for the mapping between our two approaches. This last result indicates that our moment conditions essentially match those of Kitazawa (2022) when $T = 3$. However, for $T \geq 4$, Proposition 2 imply that there are further identifying moments than those based solely on $\hbar U_{it}$ and $\hbar \Upsilon_{it}$ for the AR(1) model. Interestingly, it turns out as we demonstrate in Appendix Section B that our moment functions coincide exactly with those derived by Honoré and Weidner (2020) for the special case $T = 3$.

To the best of our knowledge, besides the AR(1) model and a few specific examples, the structure of moment conditions in models with arbitrary lag order is not fully understood in the literature. Building on Bonhomme (2012), Honoré and Weidner (2020) propose moment functions for the AR(2) model up to $T = 4$ and the AR(3) model with $T = 5$ but no results are offered beyond these special instances. Yet, this is of general interest not only to better understand the properties of DFEL models but also for practical modelling and estimation purposes. For example, Card and Hyslop (2005) argue in favor of using higher order logit specifications to better fit the behavior of a control group in the context of a welfare experiment. Relatedly, there are few results available for multivariate fixed effect models and existing methods developed for the scalar case are likely to be difficult to adapt in practice due to computational barriers. In the remaining sections, we show that our two-step approach addresses these issues by providing closed form expressions for the moment equality conditions of these more complex models.

4.4 Moment restrictions for the AR(p) logit model, $p > 1$

Allowing for more than one lag is often desirable in empirical work to model persistent stochastic processes and to better fit the data (e.g, Magnac (2000) on labour market histories, Chay et al. (1999) and Card and Hyslop (2005) on welfare recipiency). To this end, we now discuss how to extend our identification scheme to general univariate autoregressive models.

We consider

$$Y_{it} = \mathbb{1}\left\{\sum_{r=1}^{p} \gamma_{0r} Y_{it-r} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\right\}, \quad t = 1, \dots, T \tag{5}$$

for known autoregressive order $p > 1$ and vector of initial values $Y_i^0 = (Y_{i-(p-1)}, \dots, Y_{i-1}, Y_{i0})' \in \mathcal{Y}^p$, with $A_i \in \mathbb{R}$. Here, we let $\theta_0 = (\gamma_0', \beta_0')' \in \mathbb{R}^{p+K_x}$. The corresponding transition probabilities are:

$$\pi_t^{k|l_1^p}(A_i, X_i) = P(Y_{it+1} = k | Y_{it} = l_1, \dots, Y_{it-(p-1)} = l_p, X_i, A_i) = \frac{e^{k(\sum_{r=1}^p \gamma_{0r} l_r + X_{it+1}'\beta_0 + A_i)}}{1 + e^{\sum_{r=1}^p \gamma_{0r} l_r + X_{it+1}'\beta_0 + A_i}}$$

and there will be moment restrictions attached to each of the 2^p (non-redundant) transition probabilities. Before detailing the specifics of their construction, we enumerate the moment restrictions for this model as we did for the AR(1). This provides a way to ensure that we are not leaving any information on the table.

4.4.1 Impossibility results and number of moment restrictions when $p \geq 1$

Based on simulation evidence, Honoré and Weidner (2020) conjectured that AR(p) models possess $2^T - (T+p-1)2^p$ linearly independent moment conditions in panels of sufficient length. We prove this claim in Theorem 3 and establish that no moment restrictions for the common parameters exist when $T \leq p + 1$; that is with less than $2p + 1$ periods of observations per individual. To introduce the result formally, it is again convenient to consider the conditional expectation operator mapping functions of histories Y_i to their conditional expectation given $Y_i^0 = y^0, X_i = x$ and the fixed effect, i.e

$$\begin{aligned} \mathcal{E}_{y^0,x}^{(p)} : \mathbb{R}^{\mathcal{Y}^T} &\longrightarrow \mathbb{R}^{\mathbb{R}} \\ \phi(., y^0, x) &\longmapsto \mathbb{E}\left[\phi(Y_i, y^0, x) | Y_i^0 = y^0, X_i = x, A_i = .\right] \end{aligned}$$

so that for any $y \in \mathcal{Y}^T$, $\mathcal{E}_{y^0,x}^{(p)}\left[\mathbb{1}\{. = y\}\right]$ yields the conditional likelihood of history y for all possible values of A_i in the AR(p) model. That is,

$$\mathcal{E}_{y^0,x}^{(p)}\left[\mathbb{1}\{. = y\}\right] = P(Y_i = y | Y_i^0 = y^0, X_i = x, A_i = .) = a \mapsto \prod_{t=1}^{T} \frac{e^{y_t\left(\sum_{r=1}^p \gamma_{0r} y_{t-r} + x_t'\beta_0 + a\right)}}{1 + e^{\sum_{r=1}^p \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}}$$

Then the following result holds:

Theorem 3. *Consider model (5) with $T \geq 1$ and initial condition $y^0 \in \mathcal{Y}^p$. Suppose that for any $t, s \in \{1, \ldots, T-1\}$ and $y, \tilde{y} \in \mathcal{Y}^p$, $\gamma_0' y + x_t' \beta_0 \neq \gamma_0' \tilde{y} + x_s' \beta_0$ if $t \neq s$ or $y \neq \tilde{y}$. Then, the family*

$$\mathcal{F}_{y^0,p,T} = \left\{ 1, \pi_0^{y_0|y^0}(.,x), \left\{ \left(\pi_{t-1}^{y_1|y_1^{t-1},y_0,\ldots,y_{-(p-t)}}(.,x)) \right)_{y_1^{t-1} \in \mathcal{Y}^{t-1}} \right\}_{t=2}^{p}, \left\{ \left(\pi_{t-1}^{y_1|y_1^p}(.,x) \right)_{y_1^p \in \mathcal{Y}^p} \right\}_{t=p+1}^{T} \right\}$$

forms a basis of $\operatorname{Im}\left(\mathcal{E}_{y^0,x}^{(p)}\right)$ *and therefore*

1. *If* $T \leq p+1$, $\operatorname{rank}\left(\mathcal{E}_{y^0,x}^{(p)}\right) = 2^T$ *and* $\dim\left(\ker\left(\mathcal{E}_{y^0,x}^{(p)}\right)\right) = 0$
2. *If* $T \geq p+2$, $\operatorname{rank}\left(\mathcal{E}_{y^0,x}^{(p)}\right) = (T-p+1)2^p$ *and* $\dim\left(\ker\left(\mathcal{E}_{y^0,x}^{(p)}\right)\right) = 2^T - (T-p+1)2^p$

Theorem 3 generalizes Theorem 1 for AR(p) logit models with $p > 1$. It confirms the basic intuition that all the parametric content lies in the transition probabilities, no matter the lag order. Specifically, the conditional probabilities of all choice histories are spanned by the transition probabilities. In the basis $\mathcal{F}_{y^0,p,T}$, elements $\pi_0^{y_0|y^0}(.,x)$ and $\left\{ \left(\pi_{t-1}^{y_1|y_1^{t-1},y_0,\ldots,y_{-(p-t)}}(.,x)) \right)_{y_1^{t-1} \in \mathcal{Y}^{t-1}} \right\}_{t=2}^{p}$ correspond to transition probabilities that are affected by the initial condition y^0. In the AR(1) case, it reduces to $\pi_0^{y_0|y_0}(.,x)$ (see Theorem 1). The remaining basis elements are free from the initial condition and correspond to the collection of all transition probabilities in each period starting from $t = p$.

Theorem 3 is an implication of *partial fraction decompositions* and of the fact that the transition probabilities of AR(p) models admit transition functions. This property is set out in the following section. If $T \leq p+1$, $\mathcal{E}_{y^0,x}^{(p)}$ is injective and no non-trivial moment conditions can be found. Beyond this threshold, the *rank nullity theorem* which connects image and nullspace of linear maps tells us that $2^T - (T-p+1)2^p$ moment restrictions exist. Under weaker conditions on the parameters or regressors then those of the theorem, the model may admit additional moment conditions even with $T \leq p+1$.

4.4.2 Construction of transition probabilities with $p > 1$

Having clarified that $T = p+2$ is the minimum number of periods required for the existence of identifying moments, we are now ready to address the issue of their construction. The blueprint generalizes that of the AR(1) model and can be summarized as follows:

1. **Step 1)**

   (a) Start by obtaining analytical expressions of the unique transition functions for the transition probability in period $t = p$ when $T = p + 1$ [8]. Shift these expressions by one period, two periods, three periods etc to get a set of transition functions for period $t \in \{p + 1, \ldots, T - 1\}$ when $T \geq p + 2$.

   (b) Apply *partial fraction decompositions* to the expressions obtained in (a) for $t \in \{p + 1, \ldots, T - 1\}$ to generate other transition functions mapping to the same transition probabilities.

2. **Step 2)**. Take "adequate" differences of transition functions associated to the same transition probability in periods $t \in \{p + 1, \ldots, T - 1\}$ to obtain valid moments that are linearly independent.

Step 1) (a) is akin to how we started by getting closed form expressions for the transition functions in period $t = 1$ for $T = 2$ in the one lag case and then deducted a general principle for $t \geq 2$ (see Section 4). From a technical perspective, this is the only part of the two-step procedure that differs from the baseline AR(1). Indeed, **Step 2)** is fundamentally identical and **Step 1)** (b) is also unchanged for the simple reason that the transition probabilities keep the same functional form as before. That is, a logistic transformation of a linear index composed of common parameters, the regressors and the fixed effect only. Hence, the same *partial fraction expansions* apply. In light of those close similarities with the AR(1) and in order to focus on the primary issues, we defer a discussion of **Step 1)**(b) and **Step 2)** to Appendix Section C.

Theorem 4 provides the algorithm to compute the transition functions for **Step 1)** (a) for arbitrary lag order greater than one. It is based on the insight that we can leverage the transition functions of an AR($p - 1$) and *partial fraction decompositions* to generate the transition functions of an AR(p). A simple example is helpful to illustrate those ideas.

[8]The fact that the transition functions in period $t = p$ are unique when $T = p + 1$ is a direct corollary of Theorem 3. Otherwise, the difference of two distinct transition functions mapping to the same transition probability would yield a valid moment which is a contradiction.

Consider an AR(2) with $T = 3$ (i.e 5 observations in total) and suppose that we seek a transition function associated to, say, the transition probability

$$\pi_2^{0|0,1}(A_i, X_i) = \frac{1}{1 + e^{\gamma_{02} + X_{i3}'\beta_0 + A_i}}$$

The first ingredient of the theorem is to view the AR(2) model as an AR(1) model where we treat the second order lag as an additional strictly exogenous regressor. This change of perspective is advantageous since we already know how to deal with the single lag case. In particular, Lemma 2 readily gives the transition function $\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)$ for the transition probability $\pi_2^{0|0,Y_{i1}}(A_i, X_i) = P(Y_{i3} = 0|Y_{i2} = 0, Y_{i1}, X_i, A_i)$ in the sense that it verifies:

$$\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)|Y_i^0, Y_{i1}, X_i, A_i\right] = \pi_2^{0|0,Y_{i1}}(A_i, X_i)$$

This is an intermediate stage since $\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)$ does not quite map to the target of interest; indeed $\pi_2^{0|0,Y_{i1}}(A_i, X_i)$ depends on the random variable Y_{i1} unlike $\pi_2^{0|0,1}(A_i, X_i)$. To make further progress, one would intuitively need to "set" Y_{i1} to unity to make the two transition probabilities coincide. We operationalize this idea by interacting $\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)$ and Y_{i1} to achieve the desired effect in expectation:

$$\begin{aligned}\mathbb{E}\left[Y_{i1}\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)|Y_i^0, X_i, A_i\right] &= \mathbb{E}\left[Y_{i1}\pi_2^{0|0,1}(A_i, X_i)|Y_i^0, X_i, A_i\right] \\ &= \frac{1}{1 + e^{\gamma_{02} + X_{i3}'\beta + A_i}} \frac{e^{\gamma_{01}Y_{i0} + \gamma_{02}Y_{i-1} + X_{i1}'\beta_0 + A_i}}{1 + e^{\gamma_{01}Y_{i0} + \gamma_{02}Y_{i-1} + X_{i1}'\beta_0 + A_i}}\end{aligned}$$

Here, the first equality follows from the law of iterated expectations. Then, the second ingredient of the theorem is a *partial fraction expansion* (Appendix Lemma 8) to turn this product of logistic indices into $\pi_2^{0|0,1}(A_i, X_i)$. This last operation is analogous to how we constructed sequences of transition functions in the AR(1) model. It ultimately tells us that the solution is a weighted sum of $(1 - Y_{i1})$ and $Y_{i1}\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)$. Theorem 4 turns this procedure into a recursive algorithm that computes the transition functions for any lag order $p > 1$.

Theorem 4. *In model (5) with $T \geq p+1$, for all $t \in \{p, \ldots, T-1\}$ and $y_1^p \in \mathcal{Y}^p$, let*

$$k_t^{y_1|y_1^p}(\theta) = \sum_{r=1}^{p} \gamma_r y_r + X'_{it+1}\beta$$

$$k_t^{y_1|y_1^{k+1}}(\theta) = \sum_{r=1}^{k+1} \gamma_r y_r + \sum_{r=k+2}^{p} \gamma_r Y_{it-(r-1)} + X'_{it+1}\beta, \quad k = 1, \ldots, p-2, \text{ if } p > 2$$

$$u_{t-k}(\theta) = \sum_{r=1}^{p} \gamma_r Y_{it-(r+k)} + X'_{it-k}\beta, \quad k = 1, \ldots, p-1$$

$$w_t^{y_1|y_1^{k+1}}(\theta) = \left[1 - e^{(k_t^{y_1|y_1^{k+1}}(\theta) - u_{t-k}(\theta))}\right]^{y_{k+1}} \left[1 - e^{-(k_t^{y_1|y_1^{k+1}}(\theta) - u_{t-k}(\theta))}\right]^{1-y_{k+1}}, \quad k = 1, \ldots, p-1$$

and

$$\phi_\theta^{y_1|y_1^{k+1}}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i) =$$
$$\left[(1 - Y_{it-k}) + w_t^{y_1|y_1^{k+1}}(\theta)\phi_\theta^{y_1|y_1^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)Y_{it-k}\right]^{(1-y_1)y_{k+1}} \times$$
$$\left[1 - Y_{it-k} - w_t^{y_1|y_1^{k+1}}(\theta)\left(1 - \phi_\theta^{y_1|y_1^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)\right)(1 - Y_{it-k})\right]^{(1-y_1)(1-y_{k+1})} \times$$
$$\left[Y_{it-k} + w_t^{y_1|y_1^{k+1}}(\theta)\phi_\theta^{y_1|y_1^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)(1 - Y_{it-k})\right]^{y_1(1-y_{k+1})} \times$$
$$\left[1 - (1 - Y_{it-k}) - w_t^{y_1|y_1^{k+1}}(\theta)\left(1 - \phi_\theta^{y_1|y_1^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)\right)Y_{it-k}\right]^{y_1 y_{k+1}}, \quad k = 1, \ldots, p-1$$

where

$$\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) = (1 - Y_{it})e^{Y_{it+1}(\gamma_1 Y_{it-1} - \sum_{l=2}^{p}\gamma_l \Delta Y_{it+1-l} - \Delta X'_{it+1}\beta)}$$
$$\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) = Y_{it}e^{(1-Y_{it+1})\left(\gamma_1(1-Y_{it-1}) + \sum_{l=2}^{p}\gamma_l \Delta Y_{it+1-l} + \Delta X'_{it+1}\beta\right)}$$

Then,

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^p}(Y_{it+1}, Y_{it}, Y_{it-(2p-1)}^{t-1}, X_i) \,|\, Y_i^0, Y_{i1}^{t-p}, X_i, A_i\right] = \pi_t^{y_1|y_1^p}(A_i, X_i)$$

and for $k = 0, \ldots, p-2$

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{k+1}}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i) \,|\, Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] = \pi_t^{y_1|y_1^{k+1}, Y_{it-(k+1)}, \ldots, Y_{it-(p-1)}}(A_i, X_i)$$

The remaining steps to complete the construction of valid moment functions are described at length in Appendix Section C. The end product is a family of (numerically) linearly independent moment functions of size $2^T - (T+1-p)2^p$. By Theorem 3, this implies that our two-step approach recovers all *moment equality* conditions in the model.

Remark 7. (Extensions) While the exposition emphasized model (5), our methodology applies more broadly to models of the form

$$Y_{it} = \mathbb{1}\left\{g(Y_{it-1}, \ldots, Y_{it-p}, X_{it}, \theta_0) + A_i - \epsilon_{it} \geq 0\right\}, \quad t = 1, \ldots, T$$

where the lag order $p > 1$ is known and $g(.)$ is known up to the finite dimensional parameter θ_0. We can thus incorporate interaction effects which are often of interest in applied work. For instance, Card and Hyslop (2005) model welfare participation as a random effect AR(2) logit process of the form

$$Y_{it} = \mathbb{1}\left\{\gamma_{01} Y_{it-1} + \gamma_{02} Y_{it-2} + \delta_0 Y_{it-1} Y_{it-2} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\right\}, \quad t = 1, \ldots, T$$

where A_i either follows a normal distribution or a discrete distribution with few support points. In this case, minor modifications of the results in this section will deliver moment conditions for $\theta_0 = (\gamma_{01}, \gamma_{02}, \delta_0, \beta_0')'$ that are robust to misspecifications of individual unobserved heterogeneity. The key is that A_i enters additively in order to leverage the rational fraction identities of Lemma 8.

4.5 Identification with more than one lag

This section discusses ways to leverage our methodology and moment restrictions to assess the identifiability of common parameters. For ease of exposition, we concentrate on the AR(2) logit model.

We start by briefly reexamining an identification result due to Honoré and Weidner (2020). Using functional differencing, they proved (under some regularity conditions) that θ_0 is identified with $T = 3$ provided $X_{i2} = X_{i3}$ and that the initial condition $Y_i^0 = (Y_{i-1}, Y_{i0})$ varies in the population. Notice that this is not in contradiction to Theorem 3 since $X_{i2} = X_{i3}$ and Y_i^0 "varying" constitute two violations of its key assumptions. It is therefore not unsurprising that identifying moment exist in that case despite $T < 4$. To understand why, note that imposing $X_{i2} = X_{i3}$ effectively amounts to equate the transition probabilities in period $t = 2$ and in period $t = 1$ for adequate choices of the initial condition; e.g $\pi_1^{0|0,Y_{i0}}(A_i, X_i) = \pi_2^{0|0,0}(A_i, X_i)$ provided that $Y_{i0} = 0$ and $X_{i2} = X_{i3}$. In turn, this implies that differences of the corresponding transition functions in periods $t = 2$ and $t = 1$ deliver valid moment functions to estimate θ_0 in certain subpopulations. In Appendix Section J.1, we show that this is an interpretation of the moment conditions that Honoré and Weidner (2020) use to show point identification.

Because this identification argument hinges on matching covariates as in Honoré and Kyriazidou (2000), it breaks down in the presence of certain types of regressors like an age variable or a time trend. In fact, Dobronyi et al. (2021) showed that there are actually no moment equality conditions available in the model with such regressors. This finding is consistent with the intuition that we cannot match the transition probabilities in periods $t = 1$ and $t = 2$ in that case. However, with one additional period, i.e $T = 4$, we can leverage the moment restrictions of Proposition 4 which are valid for free-varying regressors and any initial condition. This leads to two possible approaches to inference. The first is to consider the "identified set" Θ^I of θ_0 based on the four conditional moment restrictions implied by the model:

$$\Theta^I = \left\{ \theta \in \mathbb{R}^{2+K_x} : \mathbb{E}_{\theta_0} \left[\psi_\theta^{y_1|y_1,y_2}(Y_{i0}^4, Y_{i-1}^1, X_i) | Y_i^0, X_i \right] = 0, \quad \forall (y_1, y_2) \in \{0,1\}^2 \right\}$$

and construct confidence sets for θ_0 following e.g Andrews and Shi (2013). Instead, the sharp identified set may be computed following the approach of Dobronyi et al. (2021) if the covariates X_i are discrete with finite support. Alternatively, a second approach which we develop further here is to formulate sensible restrictions on covariates that secure point identification in the spirit of Honoré and Kyriazidou (2000). Specifically, we consider the case where a continuous scalar component W_{i2} of X_{i2} has unbounded positive support conditional on Y_i^0, the other regressors, A_i and has a non-trivial effect β_{0W} of known sign to the econometrician. This is the content of Assumption 2 in which $Z_i = (R_i', W_{i1}, W_{i3}, W_{i4})$, and $X_{it} = (W_{it}, R_{it}') \in \mathbb{R}^{K_x}$ for all $t \in \{1, 2, 3, 4\}$. Dobronyi et al. (2023) used a similar device to develop an alternative distribution-free semiparametric estimator to that of Honoré and Kyriazidou (2000) that can accommodate time effects in the baseline one lag model.

Assumption 2. *(i) The covariate W_{i2} is continuously distributed with unbounded support* *on $\mathbb{R}_+$ conditional on Y_i^0, Z_i, A_i and (ii) β_{0W} is known to be strictly negative.*

Besides being a technical convenience, Assumption 2 may be reasonable in some situations, e.g in the context of our empirical application, the econometrician may have a confident prior that drug prices affect individual drug consumption negatively. We point out that nothing in the discussion that follows hinges critically on $\beta_W < 0$ and or W_{i2} having support on the positive reals. A set of perfectly symmetric arguments will deliver the same conclusions if instead $\beta_W > 0$ and W_{i2} has unbounded support on $\mathbb{R}_-$.

Assumption 3. *(i) $\theta_0 = (\gamma_{01}, \gamma_{02}, \beta_0')' \in \mathbb{G}_1 \times \mathbb{G}_2 \times \mathbb{B} = \Theta$, $\mathbb{G}_1, \mathbb{G}_2, \mathbb{B}$ compact. The conditional densities of A_i and Z_i verify:*

(ii) $\lim_{w_2 \to \infty} p(a|y^0, z, w_2) = q(a|y^0, z)$, $\lim_{w_2 \to \infty} p(z|y^0, w_2) = q(z|y^0)$

(iii) There exists positive integrable functions $d_0(a), d_1(z), d_2(z)$ such that $p(a|y^0, z, w_2) \leq d_0(a)$ for all $a \in \mathbb{R}$, $d_1(z) \leq p(z|y^0, w_2) \leq d_2(z)$ for all $z \in \mathbb{R}^{K_x - 1}$

(iv) $w_2 \mapsto p(a|y^0, z, w_2), w_2 \mapsto p(z|y^0, w_2)$ are continuous in w_2.

Assumption 3 are standard regularity conditions for an application of the dominated convergence theorem that once paired with Assumption 2 are sufficient to establish that θ_0 is *identified at infinity.* The outline of the argument is as follows. Under these assumptions, by sending W_{i2} to ∞, the valid moment function $\psi_\theta^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, X_i)$ of Proposition 4 reduces to

$$
\begin{aligned}
\psi_{\theta,\infty}^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, Z_i) = &-(1 - Y_{i1})(1 - Y_{i2})Y_{i3} \\
&+ \left[e^{X_{i34}'\beta} - 1\right](1 - Y_{i1})(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\
&+ e^{-\gamma_1 Y_{i0} + \gamma_2(1 - Y_{i-1}) + X_{i31}'\beta} Y_{i1}(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\
&+ e^{-\gamma_1 Y_{i0} - \gamma_2 Y_{i-1} + X_{i41}'\beta} Y_{i1}(1 - Y_{i2})(1 - Y_{i3})(1 - Y_{i4})
\end{aligned} \tag{6}
$$

which occurs because $\lim_{w_2 \to \infty} e^{w_2 \beta_W} = 0$ and $Y_{i2} = 0$ with probability one conditional on the regressors and the fixed effects. The key observation is that this "limiting" moment function has a similar functional form to the valid moment functions of the AR(1) model with $T = 3$. In turn, this implies monotonicity properties on certain regions of the covariate space that we can exploit to point identify θ_0 in the spirit of Honoré and Weidner (2020). To this end, let $(\bar{x}, \underline{x}) \in \mathbb{R}^2$, such that $\bar{x} > \underline{x}$ and define the sets

$$\mathcal{X}_{k,+} = \{x \in \mathbb{R}^{4K_x} | \bar{x} \geq x_{k,3} \geq x_{k,4} > x_{k,1} \geq \underline{x} \text{ or } \bar{x} \geq x_{k,3} > x_{k,4} \geq x_{k,1} \geq \underline{x}\}$$
$$\mathcal{X}_{k,-} = \{x \in \mathbb{R}^{4K_x} | \underline{x} \leq x_{k,3} \leq x_{k,4} < x_{k,1} \leq \bar{x} \text{ or } \underline{x} \leq x_{k,3} < x_{k,4} \leq x_{k,1} \leq \bar{x}\}$$

for all $k \in \{1, \dots, K_x\}$. In words, $\mathcal{X}_{k,+}$ is the region of the covariate space in which values of the k-th regressor in periods $t \in \{1, 3, 4\}$ belong to $[\underline{x}, \bar{x}]$ and verify $x_{k,3} \geq x_{k,4} \geq x_{k,1}$ with at least one strict inequality. Instead, $\mathcal{X}_{k,-}$ is the region of the covariate space where realizations of the k-th regressor obey the reverse ranking. With these notations in hands, we have the following theorem,

Theorem 5. *For $T = 4$, suppose that outcomes $(Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$ are generated from model (5) with $p = 2$, initial condition $y^0 \in \mathcal{Y}^2$, common parameters $\theta_0 = (\gamma_0', \beta_0') \in \mathbb{R}^{2+K_x}$ and that Assumptions 2 and 3 hold. Further, for all $s \in \{-,+\}^{K_x}$, let $\mathcal{X}_s = \bigcap_{k=1}^{K_x} \mathcal{X}_{k,s_k}$ and suppose that for all $y^0 \in \mathcal{Y}^2$*

$$\lim_{w_2 \to \infty} P\left(Y_i^0 = y^0, \quad X_i \in \mathcal{X}_s \,|\, W_{i2} = w_2\right) > 0$$

Let

$$\Psi_{s,y^0}^{0|0,0}(\theta) = \lim_{w_2 \to \infty} \mathbb{E}\left[\psi_{\theta,\infty}^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, X_i) \,|\, Y_i^0 = y^0, X_i \in \mathcal{X}_s, W_{i2} = w_2\right]$$

Then, θ_0 is the unique solution to the system of equations

$$\Psi_{s,y^0}^{0|0,0}(\theta) = 0, \quad \forall s \in \{-,+\}^{K_x}, \quad \forall y^0 \in \mathcal{Y}^2$$

Theorem 5 shows that point identification of θ_0 is achievable in higher-order dynamic logit models in short panels. The main cost for this guarantee is Assumption 2 which presumes knowledge of the data generating process beyond the baseline setup. Additionally, there should be sufficient variation in the regressors X_{it} as $W_{i2} \mapsto \infty$ to ensure that $\lim_{w_2 \to \infty} P\left(Y_i^0 = y^0, \quad X_i \in \mathcal{X}_s \,|\, W_{i2} = w_2\right) > 0$ for all $s \in \{-,+\}^{K_x}$. Our arguments are easily generalizable to AR(p) models with lag order $p \geq 3$. Under natural extensions of Assumptions 2 and 3, the model parameters $\theta_0 = (\gamma_{01}, \dots, \gamma_{0p}, \beta_0')$ are *identified at infinity* provided $T \geq 2 + p$.

Remark 8 (Identification with time effects)**.** Theorem 5 does not readily deals with time effects but it is straightforward to adapt the argument for this case. Suppose for concreteness that one covariate is a time trend. By further sending W_{i3} to infinity, the limiting moment function of equation (6) reduces to

$$\psi_{\theta,\infty}^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^{2}, Z_i) = -(1 - Y_{i1})(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\ + e^{-\gamma_1 Y_{i0} - \gamma_2 Y_{i-1} + X_{i41}'\beta} Y_{i1}(1 - Y_{i2})(1 - Y_{i3})(1 - Y_{i4})$$

For $(Y_{i0}, Y_{i-1}) = (0, 0)$, this valid moment function only depends on β and arguments analogous to those in Theorem 5 will point identify β_0. Varying the initial condition is then sufficient to point identify γ_0 given the monotonicity of the moment function in (γ_1, γ_2).

4.6 Average Marginal Effects in AR(p) logit models

In discrete choice settings, interest often centers on certain functionals of unobserved heterogeneity rather than on the value of the model parameters per se. One particular family of such functionals that are of interest from a policy perspective are average marginal effects (AMEs) which capture mean response to a counterfactual change in past outcomes. It turns out that these key quantities are simply expectations of our transition functions. To see this, consider first the baseline AR(1) model with discrete covariates X_{it}. We can define the average transition probability from state l to state k in period t for a subpopulation of individuals with covariate $x_1^{t+1} = (x_1, \ldots, x_{t+1})$ and initial condition y_0 as

$$\Pi_t^{k|l}(y_0, x_1^{t+1}) = \mathbb{E}\left[\underbrace{\pi_t^{k|l}(X_{it+1}, A_i)}_{\equiv \pi_t^{k|l}(X_i, A_i)} \,|\, Y_{i0} = y_0, X_{i1}^{t+1} = x_1^{t+1}\right] = \int \pi_t^{k|l}(x_{t+1}, a) p(a|y_0, x_1^{t+1}) da$$

where $p(a|y_0, x_1^{t+1})$ denotes the conditional density of the fixed effect A given (y_0, x_1^{t+1}). The AME is defined as the following contrast of average transition probabilities:

$$AME_t(y_0, x_1^{t+1}) = \Pi_t^{1|1}(y_0, x_1^{t+1}) - \Pi_t^{1|0}(y_0, x_1^{t+1}) = \Pi_t^{1|1}(y_0, x_1^{t+1}) - (1 - \Pi_t^{0|0}(y_0, x_1^{t+1}))$$

It is interpreted as the population average causal effect on Y_{it+1} of a change from 0 to 1 of Y_{it} given (y_0, x_1^{t+1}). By Lemma 2 and the law of iterated expectations, we have that for $T \geq 2$ and $t \geq 1$:

$$\Pi_t^{0|0}(y_0, x_1^{t+1}) = \mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) \,|\, Y_{i0} = y_0, X_{i1}^{t+1} = x_1^{t+1}\right]$$
$$\Pi_t^{1|1}(y_0, x_1^{t+1}) = \mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) \,|\, Y_{i0} = y_0, X_{i1}^{t+1} = x_1^{t+1}\right]$$

which implies that $AME_t(y_0, x_1^{t+1})$ is identified so long as θ_0 is identified. A sufficient condition for that is $T \geq 3$ and $X_{i3} - X_{i2}$ having support in a neighborhood of zero (Honoré and Kyriazidou (2000)). Aguirregabiria and Carro (2021) were the first to highlight that AMEs can be point identified in the AR(1) model. When the lag order p is greater than one - which seems to be the case for persistent variables such as unemployment (e.g Magnac (2000)) and welfare recipiency (e.g Chay et al. (1999)) - we can analogously define average transition probabilities from states $l_1^p \in \mathcal{Y}^p$ to state $k \in \mathcal{Y}$ as:

$$\Pi_t^{k|l_1^p}(y^0, x_1^{t+1}) = \mathbb{E}\left[\underbrace{\pi_t^{k|l_1^p}(X_{it+1}, A_i)}_{\equiv \pi_t^{k|l}(X_i, A_i)} \,|\, Y_i^0 = y^0, X_{i1}^{t+1} = x_1^{t+1}\right] = \int \pi_t^{k|l_1^p}(x_{t+1}, a) p(a|y_0, x_1^{t+1}) da$$

This permits the consideration of more nuanced counterfactual parameters compared to the AR(1). In the context of studies on long term unemployment, contrasts of the form $\Pi_t^{k|l_1^p}(y^0, x_1^{t+1}) - \Pi_t^{k|v_1^p}(y^0, x_1^{t+1})$ may be especially relevant to measure more accurately the relative effects of work histories spanning multiple periods. Again, these counterfactuals are simply expectations of transition functions by Theorem 4 and will be identified whenever θ_0 is identified (see Section 4.5 for examples of sufficient conditions).

Multiperiod analogs of average transition probabilities in AR(p) models

$$\Pi_t^{k_1^s|l_1^p}(y^0, x_1^{t+s}) =$$
$$\mathbb{E}\left[P(Y_{it+s} = k_s, \ldots, Y_{it+1} = k_1 \,|\, Y_{it} = l_1, \ldots, Y_{it-(p-1)} = l_p, X_{i1}^{t+s} = x_1^{t+s}, A_i) \,|\, Y_i^0 = y^0, X_{i1}^{t+s} = x_1^{t+s}\right]$$

may also be of interest to assess state-dependence. These quantities give the average probability of moving from states $l_1^p \in \mathcal{Y}^p$ to future states $k_1^s \in \mathcal{Y}^s$, where $s \geq 1$ and the average is taken with respect to the distribution of A_i conditional on (y_0, x_1^{t+1}). The special case

$k_1 = k_2 = \ldots = k_s$ delivers a discrete version of the survivor function employed in duration analysis, i.e the average likelihood to survive s consecutive periods in the same state after experiencing a given choice history. Proposition 3 shows that they are also identified when θ_0 is identified under certain conditions.

Proposition 3. *Consider model (5) with $T \geq p+2$, and initial condition $y^0 \in \mathcal{Y}^p$. Suppose that θ_0 is identified and that for any $t \in \{p, \ldots, T-2\}$, $s \in \{1, \ldots, T-1-t\}$ and $y, \tilde{y} \in \mathcal{Y}^p$, $\gamma_0' y + x_t'\beta_0 \neq \gamma_0'\tilde{y} + x_{t+s}'\beta_0$. Then, for $t \in \{p, \ldots, T-2\}$, $s \in \{1, \ldots, T-1-t\}$, and any $l_1^p \in \mathcal{Y}^p$, $k_1^s \in \mathcal{Y}^s$, the quantity $\Pi_t^{k_1^s|l_1^p}(y^0, x_1^{t+s})$ is identified.*

The source of this result is the fact that the integrand of $\Pi_t^{k_1^s|l_1^p}(y^0, x_1^{t+s})$ is a product of transition probabilities. This entails that under appropriate conditions on the regressors and common parameters, we can turn this integrand into a unique linear combination of transition probabilities by means of a *partial fraction decomposition*. It is then a matter of taking expectations and invoking the fact that average transition probabilities are identified from our transition functions.

Example 1 (Survivor function for an AR(2))**.** To illustrate Proposition 3, and in the spirit of our upcoming empirical application, suppose that Y_{it} is an indicator for drug consumption at time t obeying an AR(2) logit process. Fix $y^0 \in \mathcal{Y}^2$ and assume $T = 5$. One might be interested in

$$
\begin{aligned}
\Pi_3^{0,0|1,1}(y^0, x) &= \mathbb{E}\left[P(Y_{i5} = 0, Y_{i4} = 0 \,|\, Y_{i3} = 1, Y_{i2} = 1, X_i = x, A_i) \,|\, Y_i^0 = y^0, X_i = x\right] \\
&= \mathbb{E}\left[\pi_4^{0|0,1}(A_i, x)\pi_3^{0|1,1}(A_i, x) \,|\, Y_i^0 = y^0, X_i = x\right]
\end{aligned}
$$

which gives the average propensity of individuals with characteristics (y^0, x) who consumed drugs in $t = 2, 3$ to stay drug-free over the next two time periods. A simple calculation using for instance the identities of Appendix Lemma 8 gives

$$
\begin{aligned}
\pi_4^{0|0,1}(A_i, x)\pi_3^{0|1,1}(A_i, x) &= \frac{1}{1 + e^{\gamma_{02} + x_5'\beta_0 + A_i}} \frac{1}{1 + e^{\gamma_{01} + \gamma_{02} + x_4'\beta_0 + A_i}} \\
&= \frac{1}{1 - e^{\gamma_{01} + x_{45}'\beta_0}} \pi_4^{0|0,1}(A_i, x) - \frac{e^{\gamma_{01} + x_{45}'\beta_0}}{1 - e^{\gamma_{01} + x_{45}'\beta_0}} \pi_3^{0|1,1}(A_i, x)
\end{aligned}
$$

and since Theorem 4 implies $\mathbb{E}\left[\phi_{\theta_0}^{0|0,1}(Y_{i1}^5, x) \,|\, Y_i^0 = y^0, Y_{i1}^2, X_i = x, A_i\right] = \pi_4^{0|0,1}(A_i, x)$ and

$\mathbb{E}\left[\phi_{\theta_0}^{0|1,1}(Y_{i0}^4, x)\,|\,Y_i^0 = y^0, Y_{i1}, X_i = x, A_i\right] = \pi_3^{0|0,1}(A_i, x)$, we obtain

$$\Pi_3^{0,0|1,1}(y^0, x) = \mathbb{E}\left[\frac{1}{1 - e^{\gamma_{01} + x_{45}'\beta_0}}\phi_{\theta_0}^{0|0,1}(Y_{i1}^5, x) - \frac{e^{\gamma_{01} + x_{45}'\beta_0}}{1 - e^{\gamma_{01} + x_{45}'\beta_0}}\phi_{\theta_0}^{0|1,1}(Y_{i0}^4, x)\,|\,Y_i^0 = y^0, X_i = x\right]$$

5 Multi-dimensional fixed effects models

We now turn our attention to multi-dimensional fixed effects models. We show that the general blueprint developed in the scalar case to derive valid moment functions carries over to VAR(1) and MAR(1) models. We make no attempt at showing that our approach is exhaustive in those cases and do not claim that it is. We leave these important questions for future work. Readers uninterested in the details of the multivariate extensions can skip directly to Section 6 where we discuss the empirical application.

5.1 Moment restrictions for the VAR(1) logit model

We begin with the analysis of VAR(1) logit models, variants of which have been successfully used to study the relationship between sickness and unemployment (Narendranthan et al. (1985)), the progression from softer drug use to harder drug use among teenagers (Deza (2015)), transitivity in networks (Graham (2013), Graham (2016)) and more recently the employment of couples (Honoré et al. (2022)). For a given $M \geq 2$, the model reads:

$$Y_{m,it} = \mathbb{1}\left\{\sum_{j=1}^{M}\gamma_{0mj}Y_{j,it-1} + X_{m,it}'\beta_{0m} + A_{m,i} - \epsilon_{m,it} \geq 0\right\}, \quad m = 1, \ldots, M, \quad t = 1, \ldots, T \tag{7}$$

We let $Y_{it} = (Y_{1,it}, \ldots, Y_{M,it})'$ denote the outcome vector in period t with support $\mathcal{Y} = \{0,1\}^M$ of cardinality 2^M. We let $X_{it} = (X_{1,it}', \ldots, X_{M,it}')' \in \mathbb{R}^{K_1} \times \ldots \times \mathbb{R}^{K_M}$ denote the vector of exogenous covariates in period t and $A_i = (A_{1,i}, \ldots, A_{M,i})' \in \mathbb{R}^M$. The initial condition is now given by $Y_{i0} = (Y_{1,i0}, \ldots, Y_{M,i0})' \in \mathcal{Y}$ and the model transition probabilities are given by:

$$\pi_t^{k|l}(A_i, X_i) = P(Y_{it+1} = k|Y_{it} = l, X_i, A_i) = \prod_{m=1}^{M}\frac{e^{k_m(\sum_{j=1}^{M}\gamma_{0mj}l_j + X_{m,it+1}'\beta_{0m} + A_{m,i})}}{1 + e^{\sum_{j=1}^{M}\gamma_{0mj}l_j + X_{m,it+1}'\beta_{0m} + A_{m,i}}}$$

for all $(k,l) \in \mathcal{Y} \times \mathcal{Y}$.

Building on Honoré and Kyriazidou (2000), Honoré and Kyriazidou (2019) use a conditional likelihood approach to prove the identification $\theta_0 = (\gamma_{011}, \gamma_{012}, \gamma_{021}, \gamma_{022}, \beta_{01}, \beta_{02})$ for the bivariate specification when $T = 3$ and the regressors do not vary over the last two periods. As in scalar models, we show hereinafter that this strong restriction which can yield undesirable rates of convergence is unnecessary to obtain valid moment conditions.

Step 1) in the VAR(1) logit model has a nuance relative to its scalar counterpart in that the only transition functions that appear to exist are those associated to $\pi_t^{k|k}(A_i, X_i)$, for $k \in \mathcal{Y}$, i.e the probabilities of staying in the same state. We can use the same heuristic as in the baseline AR(1) model to derive their expressions, especially in the bivariate case. Once all four transition functions are obtained for the case $M = 2$, it becomes clear that the general functional form is as per Lemma 4. It is then a matter of brute force calculation to verify that this is indeed correct.

Lemma 4. *In model (7) with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, let for all $k \in \mathcal{Y}$*

$$\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = \mathbb{1}\{Y_{it} = k\} e^{\sum_{m=1}^{M}(Y_{m,it+1}-k_m)\left(\sum_{j=1}^{M}\gamma_{mj}(Y_{j,it-1}-k_j) - \Delta X'_{m,it+1}\beta_m\right)}$$

Then:

$$\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i) = \prod_{m=1}^{M} \frac{e^{k_m(\sum_{j=1}^{M}\gamma_{0mj}k_j + X'_{m,it+1}\beta_{0m} + A_{m,i})}}{1 + e^{\sum_{j=1}^{M}\gamma_{0mj}k_j + X'_{m,it+1}\beta_{0m} + A_{m,i}}}$$

Next, we can appeal to the second *partial fraction decomposition* formula in Appendix Lemma 9 to guide the construction of another set of transition functions when $T \geq 3$. These identities may be regarded as a generalization of Kitazawa (2022)'s hyperbolic transformations to the multivariate case. As is clear from Lemma 5, the resulting transition functions have a special structure that generalizes those found in the AR(1) model.

Lemma 5. *In model (7) with $T \geq 3$, for all t, s such that $T - 1 \geq t > s \geq 1$, let for all*

$m \in \{1, \ldots, M\}$ and $(k,l) \in \mathcal{Y}^2$

$$\mu_{m,s}(\theta) = \sum_{j=1}^{M} \gamma_{mj} Y_{j,is-1} + X'_{m,is}\beta_m$$
$$\kappa^{k|k}_{m,t}(\theta) = \sum_{j=1}^{M} \gamma_{mj} k_j + X'_{m,it+1}\beta_m$$
$$\omega^{k|k}_{t,s,l}(\theta) = 1 - e^{\sum_{j=1}^{M}(l_j - k_j)\left[\kappa^{k|k}_{j,t}(\theta) - \mu_{j,s}(\theta)\right]}$$

and define the moment functions

$$\zeta^{k|k}_{\theta}(Y^{t+1}_{it-1}, Y^{s}_{is-1}, X_i) = \mathbb{1}\{Y_{is} = k\} + \sum_{l \in \mathcal{Y}\setminus\{k\}} \omega^{k|k}_{t,s,l}(\theta)\mathbb{1}\{Y_{is} = l\}\phi^{k|k}_{\theta}(Y^{t+1}_{it-1}, X_i)$$

Then,

$$\mathbb{E}\left[\zeta^{k|k}_{\theta_0}(Y^{t+1}_{it-1}, Y^{s}_{is-1}, X_i)|Y_{i0}, Y^{s-1}_{i1}, X_i, A_i\right] = \pi^{k|k}_{t}(A_i, X_i)$$

Beyond $T = 4$, more transition functions are available and can be derived sequentially from those of Lemma 5. See Corollary 5.1 for their expressions.

Corollary 5.1. *In model (7) with $T \geq 4$, for any t and ordered collection of indices s^J_1, $J \geq 2$, satisfying $T - 1 \geq t > s_1 > \ldots > s_J \geq 1$, let for all $k \in \mathcal{Y}$*

$$\zeta^{k|k}_{\theta}(Y^{t+1}_{it-1}, Y^{s_1}_{is_1-1}, \ldots, Y^{s_J}_{is_J-1}, X_i) = \mathbb{1}\{Y_{is_J} = k\} + \sum_{l \in \mathcal{Y}\setminus\{k\}} \omega^{k|k}_{t,s_J,l}(\theta)\mathbb{1}\{Y_{is_J} = l\}\zeta^{k|k}_{\theta}(Y^{t+1}_{it-1}, Y^{s_1}_{is_1-1}, \ldots, Y^{s_{J-1}}_{is_{J-1}-1}, X_i)$$

with weights $\omega^{k|k}_{t,s_J,l}(\theta)$ defined as in Lemma 5. Then,

$$\mathbb{E}\left[\zeta^{k|k}_{\theta_0}(Y^{t+1}_{it-1}, Y^{s_1}_{is_1-1}, \ldots, Y^{s_J}_{is_J-1}, X_i)|Y_{i0}, Y^{s_J-1}_{i1}, X_i, A_i\right] = \pi^{k|k}_{t}(A_i, X_i)$$

Step 2). One can obtain a family of valid moment functions by adequately repurposing the statement of Proposition 2 to the VAR(1) case, i.e by updating the expressions of $\phi^{k|k}_{\theta}(.)$ and $\zeta^{k|k}_{\theta}$ according to Lemma 4 and Corollary 5.1. To conserve on space and avoid repetition, we leave this simple exercise to the reader.

Remark 9 (Network Extension)**.** Similarly to Remarks 7, we emphasize that the tools developed here can be modified to handle other interesting variants featuring more complex interdependencies across the different layers of the model indexed by $m = 1, \dots, M$. To illustrate the wider applicability of our two-step method, we show in Appendix N how one can derive moment restrictions in the dynamic network formation model of Graham (2013) and extensions thereof incorporating exogenous covariates.

5.2 Moment restrictions for the dynamic multinomial logit model

Last, we cover dynamic multinomial logit models which have been utilized to measure state-dependence in a range of economic contexts including: employment history in the French labor market (Magnac (2000)), the impact of international trade on the transition matrix of employment across sectors (Egger et al. (2003)) and consumer product choice (Dubé et al. (2010)) amongst others.

We focus on the the baseline MAR(1) logit model with fixed effects.

The model assumes a fixed number of alternatives $C+1$ with $C \geq 1$ and is characterized by the following transition probabilities:

$$\pi_t^{k|l}(A_i, X_i) = P(Y_{it+1} = k | Y_{it} = l, X_i, A_i) = \frac{e^{\gamma_{kl} + X'_{ikt+1}\beta_k + A_{ik}}}{\sum\limits_{c=0}^{C} e^{\gamma_{cl} + X'_{ict+1}\beta_j + A_{ic}}}, \quad t = 1, \dots, T \tag{8}$$

with $(k, l) \in \mathcal{Y} = \{0, 1, \dots, C\}$. Here, $Y_{it} \in \mathcal{Y}$ indicates the choice of individual i in period t, X_{ijt} denotes a vector of individual-alternative specific exogenous covariates and $A_{ij} \in \mathbb{R}$ is the fixed effect attached to alternative j for individual i. The initial condition is $Y_{i0} \in \mathcal{Y}$ and in keeping with the fixed effect assumption, its conditional distribution given unobserved heterogeneity and the regressors, $\big(P(Y_{i0} = k | X_i, A_i)\big)_{k=1}^{C}$, is left fully unrestricted. Following Magnac (2000), we normalize the transition parameters and fixed effect of the reference alternative "0" to zero [9]. That is $\gamma_{j0} = \gamma_{0j} = 0$ for all $j \in \mathcal{Y}$ and $A_{i0} = 0$, leaving $\theta = \big((\gamma_{kl})_{k,l \geq 1}, (\beta_l)_{l \geq 0}\big)$ as the unknown model parameters.

This specification can be motivated by assuming that agents rank options according to random latent utility indices with disturbances independent over time and across alterna-

[9]The transition parameters of the reference state cannot be identified so a normalization constraint must be imposed. Setting $A_{i0} = 0$ is also without loss of generality since we can always redefine the fixed effect as $A^*_{ik} = A_{ik} - A_{i0}$.

tives. In this context, equation (8) is obtained if the best alternative is selected and the error terms are Type 1 extreme value distributed conditional on Y_{i0}, A_i, X_i. Magnac (2000) studies the "pure" case without covariates and shows that an extension of the conditional likelihood approach proposed by Chamberlain (1985) can be used to identify and estimate the state-dependence parameters. Honoré and Kyriazidou (2000) show that this argument carries over to the case with exogenous explanatory variables if one matches the regressors across specific time periods. Here, we offer an alternative estimation strategy that circumvents the need for matching.

Step 1). Similarly to the VAR(1) model the MAR(1) appears to admit transition functions only for the probabilities of staying in the same state, namely $\pi_t^{k|k}(A_i, X_i)$ for $k \in \mathcal{Y}$. This feature appears to be a common trait of multidimensional fixed effects specifications. To facilitate the derivation of the relevant transition functions, we follow our usual heuristic of looking for $\phi_\theta^{k|k}(.), k \in \mathcal{Y}$ satisfying:

$$\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = \mathbb{1}\{Y_{it} = k\}\phi_\theta^{k|k}(Y_{it+1}, k, Y_{it-1})$$
$$\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) \mid Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i)$$

Upon obtaining their exact expressions for the simplest case with $C = 2$, it is easy to conjecture and verify by direct calculations that the general expressions of the $C+1$ transition functions of the MAR(1) model are as displayed in Lemma 6.

Lemma 6. *In model (8) with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, let for all $k \in \mathcal{Y}$*

$$\phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i) = \mathbb{1}\{Y_{it} = k\}e^{\sum_{c \in \mathcal{Y}\setminus\{k\}} \mathbb{1}\{Y_{it+1}=c\}\left(\sum_{j\in\mathcal{Y}}(\gamma_{cj}-\gamma_{kj})\mathbb{1}(Y_{it-1}=j)+\gamma_{kk}-\gamma_{ck}+\Delta X'_{ikt+1}\beta_k-\Delta X'_{ict+1}\beta_c\right)}$$

Then:

$$\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i) == \frac{e^{\gamma_{kk}+X'_{ikt+1}\beta_k+A_{ik}}}{\sum\limits_{c=0}^{C} e^{\gamma_{ck}+X'_{ict+1}\beta_j+A_{ic}}}$$

Unsurprisingly, given the similarities shared between the MAR(1) and all other specifications discussed in the paper, so long as $T \geq 3$, one can again derive transition functions other than $\phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i)$ also associated to $\pi_t^{k|k}(A_i, X_i)$ for $k \in \mathcal{Y}$ in periods $t \in \{1, \ldots, T-1\}$. The simple logistic identities of Appendix Lemma 8 imply that these transition functions, that we keep denoting $\zeta_\theta^{k|k}(.)$ have a similar form to those of the VAR(1) model as shown in Lemma 7.

Lemma 7. *In model (8) with $T \geq 3$, for all t, s such that $T-1 \geq t > s \geq 1$, let for all $(c,k) \in \mathcal{Y}^2$*

$$\begin{aligned}
\mu_{c,s}(\theta) &= \sum_{j=1}^{C} \gamma_{cj} \mathbb{1}(Y_{is-1} = j) + X'_{ics}\beta_c - X'_{i0s}\beta_0 \\
\kappa_{c,t}^{k|k}(\theta) &= \gamma_{ck} + X'_{ict+1}\beta_c - X'_{i0t+1}\beta_0 \\
\omega_{t,s,c}^{k|k}(\theta) &= 1 - e^{(\kappa_{c,t}^{k|k}(\theta) - \mu_{c,s}(\theta)) - (\kappa_{k,t}^{k|k}(\theta) - \mu_{k,s}(\theta))}
\end{aligned}$$

and define the moment functions

$$\zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) = \mathbb{1}\{Y_{is} = k\} + \sum_{l \in \mathcal{Y} \setminus \{k\}} \omega_{t,s,l}^{k|k}(\theta) \mathbb{1}\{Y_{is} = l\} \phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i)$$

Then,

$$\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) | Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i)$$

Additionally, if the econometrician has access to a dataset with more than four observations per sampling unit - counting the initial condition - then, more transition functions associated to the same transition probabilities are available per Corollary 7.1.

Corollary 7.1. *In model (8) with $T \geq 4$, for any t and ordered collection of indices s_1^J, $J \geq 2$, satisfying $T-1 \geq t > s_1 > \ldots > s_J \geq 1$, let for all $k \in \mathcal{Y}$*

$$\begin{aligned}
\zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) &= \mathbb{1}\{Y_{is_J} = k\} \\
&\quad + \sum_{l \in \mathcal{Y} \setminus \{k\}} \omega_{t,s_J,l}^{k|k}(\theta) \mathbb{1}\{Y_{is_J} = l\} \zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_{j-1}-1}^{s_{J-1}}, X_i)
\end{aligned}$$

with weigts $\omega_{t,s_J,l}^{k|k}(\theta)$ defined as in Lemma 7. Then,

$$\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) | Y_{i0}, Y_{i1}^{s_J-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i)$$

This completes **Step 1)** for the MAR(1) logit model. For **Step 2)**, we recommend a family of valid moment functions mirroring those of Proposition 2 for the AR(1) case to ensure the linear independence of its elements.

6 Empirical Illustration

In this last section, we illustrate the usefulness of our methodology by revisiting the analysis of Deza (2015) on the dynamics of drug consumption amongst young adults in the United States.[10]

To provide context, multiple studies have documented that young individuals who experiment with soft drugs have a tendency to continue using them and are at a higher risk of transitioning to hard drugs. Such correlations are certainly concerning. However, the empirical evidence of genuine causal links, in particular from softer drugs to harder drugs, remains limited with Deza (2015) standing as a notable exception. Fundamentally, these empirical regularities may be attributed to a causal effect (i.e. state dependence within and between drugs) or alternatively to latent traits that make individuals more prone to using illicit substances in general. Our primary concern is to untangle these two explanations to inform the design of policies aiming to mitigate drug addiction [11]. For example, if marijuana consumption indeed serves as a gateway to later cocaine use, early educational interventions cautioning against casual marijuana usage could potentially have enduring effects on the population of heavy drug users.

To investigate these issues, we employ the restricted version of the National Longitudinal Survey of Youth 1997 (NLSY97). This is a panel dataset of 8984 individuals surveyed on a diverse range of subjects, including drug-related matters from 1997 to 2019. We concentrate on a subsample of four waves, spanning from 2001 to 2004. This subsample provides insight into the behavior of young adults between the age of 16 and 22 in 2001 to 19 and 25 in 2004. We shall examine the statistical association between three binary outcome variables, namely the consumption of alcohol, marijuana and hard drugs, derived from respondents answers' during annual interviews. Upon retaining those providing answers in all four waves as well

[10] This research was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed here are those of the author and do not reflect the views of the BLS.

[11] See Heckman (1981) for insights on the implications of state dependence for the design of labor market policies.

as a valid state of residence, our cross section ultimately consists of $N = 6317$ individuals [12]. Following Deza (2015), we then consider the trivariate VAR(1) logit model

$$Y_{m,it} = \mathbb{1}\left\{\sum_{j=1}^{3} \gamma_{0mj} Y_{j,it-1} + \beta_{0m} age_{it} + \rho_{0m} TEDS_{m,it} + \nu_{01}\mathbb{1}\{age_{it} \geq 21\}\mathbb{1}\{m = 1\} + A_{m,i} - \epsilon_{m,it} \geq 0\right\}$$

$m \in \{1, 2, 3\}$ (1="alcohol", 2="marijuana", 3="hard drugs"), $t = 1, 2, 3$ where $t = 0$ corresponds to the year 2001. The state-dependence coefficients γ_{0mm} (within) and $\gamma_{0mj}, m \neq j$ (between) are the principal coefficients of interest in the 16-dimensional vector of common parameters θ_0. We are most particularly concerned about the sign and the statistical significance of γ_{032}, i.e the so called "stepping-stone" effect of marijuana on hard drugs. The covariate age_{it} denotes the age of respondent i at time t. The regressors $TEDS_{m,it}$ measure state-level deviations from national trends in treatment admissions for substance abuse caused by drug m in year t in the state of residence of i[13]. They are computed as the ratio of the share of admissions to treatment centers due to drug m in the state of i in year t against the country wide analog in year t. Intuitively, this may be interpreted as a measure of exposure to substance m for each respondent in our sample.

Deza (2015) parameterizes both the latent permanent heterogeneity $(A_{m,i})_{m=1}^{3}$ and the initial condition Y_i^0 to estimate the model by maximum likelihood. We leave these components unrestricted and exploit the valid moment functions presented in Section 5.1. We specifically use six of the eight valid moment functions available: $\psi_\theta^{k|k}(Y_{i1}^3, Y_{i0}^1, X_i)$ for $k \in \{(0,0,0), (0,1,0), (1,1,1), (1,1,0), (1,0,1), (1,0,0)\}$. The other two corresponding to states $k \in \{(0,0,1), (0,1,1)\}$ are null for over 99.5% of our sample and were dropped to mitigate noise in estimation. Next, we (arbitrarily) select a constant, the initial condition Y_i^0, age_{it} and the covariates $TEDS_{m,it}$ in all periods $t = 1, 2, 3$ as instruments to form the 96×1 moment vector

[12] We adapt the sample selection procedure described in Deza (2015) for the period 2001-2004.

[13] The variables $TEDS_{m,it}$ are constructed from the Treatment Episode Data Set-Admissions which records admissions to substance abuse treatment facilities in the United States.

$$m_\theta(Y_i, Y_i^0, X_i) = \begin{pmatrix} \psi_\theta^{(0,0,0)|(0,0,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(0,1,0)|(0,1,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,1,1)|(1,1,1)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,1,0)|(1,1,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,0,1)|(1,0,1)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,0,0)|(1,0,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \end{pmatrix} \otimes \begin{pmatrix} 1 \\ Y_i^{0'} \\ age_{i1}^{3'} \\ TEDS_{1,i1}^{3'} \\ TEDS_{2,i1}^{3'} \\ TEDS_{3,i1}^{3'} \end{pmatrix}$$

With $m_\theta(Y_i, Y_i^0, X_i)$ in hand, we then consider the iterated GMM estimator of Hansen et al. (1996). Starting from an initial candidate $\hat{\theta}_0$[14], it can be described as

$$\hat{\theta} = \lim_{s \to \infty} \hat{\theta}_s$$
$$\hat{\theta}_s = \arg\min_\theta \overline{m}_N(\theta)' \overline{W}_N(\hat{\theta}_{s-1})^{-1} \overline{m}_N(\theta)$$

where $\overline{m}_N(\theta) = \frac{1}{N}\sum_{i=1}^N m_\theta(Y_i, Y_i^0, X_i)$ and $\overline{W}_N(\theta) = \frac{1}{N}\sum_{i=1}^N m_\theta(Y_i, Y_i^0, X_i) m_\theta(Y_i, Y_i^0, X_i)'$. Under some regularity conditions (Hansen and Lee (2021)), this estimator is well defined and asymptotically normally distributed with

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (M_0' W_0^{-1} M_0)^{-1})$$

where $M_0 = \mathbb{E}\left[\frac{\partial m_{\theta_0}(Y_i, Y_i^0, X_i)}{\partial \theta}\right]$ and $W_0 = \mathbb{E}\left[m_{\theta_0}(Y_i, Y_i^0, X_i) m_{\theta_0}(Y_i, Y_i^0, X_i)'\right]$. Our motivation for focusing on this specific estimator originates mainly from Hansen and Lee (2021) who advocate its use for two practical reasons. First, for a given set of moments, it eliminates the arbitrariness in the choice of the initial weight matrix of 2-step GMM estimators (see also Imbens (2002)). Second, because the iteration sequence is a contraction, each iteration is approximately variance reducing in the sense that: $Var(\hat{\theta}_s) \approx c^2 Var(\hat{\theta}_{s-1})$ for some constant $c < 1$ [15]. Empirically, we also found in Monte Carlo simulations that the iterated GMM estimator performs relatively well for this type of specification (see Appendix D).

Table 1 presents the iterated GMM estimates for the trivariate VAR(1) logit model in

[14]In practice, we used the GMM estimator putting equal weights on each moment as our starting candidate.

[15]Note that the limiting variance of the iterated GMM estimator and a 2-step GMM estimator will be identical.

columns (I), (II), (III). For comparison, columns (IV), (V), (VI) report a random effect (RE) estimator akin to Deza (2015) [16] while columns (VII), (VIII), (IV) display the "naive" logit maximum likelihood estimator (MLE) neglecting the presence of fixed effects.

The first observation is that, in line with conventional wisdom, GMM estimates for the state-dependence parameters within drug, $\gamma_{11}, \gamma_{22}, \gamma_{33}$, are all positive. As is apparent from columns (I)-(III), they are statistically significant for alcohol and marijuana but surprisingly not for hard drugs. In other words, there is no statistical evidence of a direct effect from past consumption of hard drug to future usage of hard drugs once we account for unobserved heterogeneity and the effects of other substances, at least in our four-wave sample[17]. Notice that the magnitude of the estimates for $\gamma_{11}, \gamma_{22}, \gamma_{33}$ sharply contrast with the other two estimators. The naive MLE largely overestimates the amount of within state-dependence, yielding coefficients that are comparatively four to eight times larger. Intuitively, this can be rationalized by the fact that this estimator misinterprets any serial correlation produced by A_i as evidence of state dependence. The RE estimator borrowed from Deza (2015) (see also Card and Hyslop (2005), Chay and Hyslop (1998)) acts as an intermediate estimator between the other two as can be seen in columns (IV)-(VI). This behavior is expected to the extent that the additional parametric structure of this methodology will account to some degree for the presence of unobserved heterogeneity. We note that the role of within state dependence in the dynamics of drug consumption is nevertheless overstated by this approach.

Second and importantly, we observe in column (III) a positive and statistically significant effect of marijuana on hard drugs. This supports the view that marijuana usage can be a gateway to the consumption of harder drugs and accords with the key findings of Deza (2015). From a practical standpoint, this result corroborates that there may be scope for policies on marijuana usage to indirectly curb the consumption of more lethal substances by teenagers and young adults. The efficacy of such policies in the short and long run are important questions that will intuitively depend on the distribution of heterogeneity in the population.

[16]We borrow the specification presented in Deza (2015). The heterogeneity distribution is discrete with 3 mass points and is independent of the regressors. The initial condition relates to the covariates through a logistic regression.

[17]The transition parameters for hard drugs are expected to be noisier given that a smaller fraction of individuals consume these more lethal substances: approximately 15% of the respondents indicate having consumed hard drugs at least once from 2001-2004. This contrasts with 86% for alcohol and 40% for marijuana.

We do not explore those questions here but further research in this direction would be of interest [18]. The other two estimators also agree on a positive influence of marijuana on the consumption of harder drugs, albeit it is statistically insignificant in the RE case.

Table 1: Parameter estimates of the trivariate VAR(1) logit

	Iterated GMM			Random Effects			Naive MLE		
	A	M	HD	A	M	HD	A	M	HD
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IV)
γ_{m1}	**0.30**	-0.04	-0.02	**1.41**	-0.36	-0.2	**2.44**	0.87	0.77
	(0.12)	(0.21)	(0.32)	(0.16)	(0.22)	(0.63)	(0.06)	(0.14)	(0.37)
γ_{m2}	-0.07	**0.70**	**0.69**	-0.52	**1.48**	**0.16**	0.72	**2.55**	**1.43**
	(0.16)	(0.14)	(0.22)	(0.12)	(0.13)	(0.25)	(0.07)	(0.07)	(0.16)
γ_{m3}	-0.20	0.26	**0.32**	-0.66	-0.17	**1.59**	0.22	0.74	**2.12**
	(0.27)	(0.22)	(0.21)	(0.19)	(0.13)	(0.13)	(0.12)	(0.09)	(0.12)
age	0.06	-0.18	0.08	0.04	-0.14	-0.05	-0.08	-0.13	-0.21
	(0.05)	(0.06)	(0.09)	(0.6)	(0.27)	(0.32)	(0.03)	(0.02)	(0.03)
age $\geq$ 21	0.04			0.46			0.54		
	(0.11)			(0.2)			(0.07)		
$TEDS_1$	-0.09			0.96			0.67		
	(0.09)			(0.77)			(0.50)		
$TEDS_2$		-0.18			0.02			-0.13	
		(0.12)			(0.48)			(0.30)	
$TEDS_3$			0.42			0.15			-0.10
			(0.32)			(0.44)			(0.40)
N		6317			6317			6317	
Periods		2001-2004			2001-2004			2001-2004	
# Iterations		12							

NOTES: *The convergence criterion of our iterated GMM procedure is* $\left\|\hat{\theta}_{s+1} - \hat{\theta}_s\right\| < 10^{-4}$. *Estimated standard errors are reported in parenthesis.*

Otherwise, it is noteworthy that the between state dependence estimates can vary quite significantly across specifications. Again, the naive MLE likely misinterprets spurious correlation

[18] A natural idea to gauge the effectiveness of policy interventions would be to compute average marginal effects. However, as mentioned in Section 5.1, we were unable to find transition functions for the transition probabilities where the state switches in VAR(1) models. This leads us to believe that only the average transition probabilities where the state remains unchanged are identified. In turn, this would imply that average marginal effects are generally partially identified in VAR(1) models. In this case, it is possible that ideas analogous to those in Dobronyi et al. (2021) and Davezies et al. (2021) could be used to characterize and compute the identified set of average marginal effects; albeit some difficulties might arise due to the fact that the fixed effects are now multidimensional. Computing outer bounds as in Pakel and Weidner (2023) could be another plausible option.

from the A_i as state dependence which results in positive and inflated cross effects. Column (IV) and (I) show disagreements of the RE and GMM estimates regarding the strength of the impact of marijuana and hard drugs on alcohol. Overall, this comparative exercise has showed that accounting for unobserved heterogeneity as flexibly as possible can be essential to obtain an accurate picture of the patterns of state dependence in practice.

7 Conclusion

Dynamic discrete choice models are widely used to study the determinants of repeated decisions made by individuals or firms over time. In this paper, we have introduced a procedure to estimate a family of such models with logistic (or Type I extreme value) errors and potentially many lags while remaining agnostic about the nature of unobserved individual heterogeneity. This type of approach may be attractive when the risk of misspecifying the initial condition and the unit-specific effects are important. We also provided general expressions for average marginal effects in the binary response case which are often the counterfactuals of interest in practice.

The list of discrete choice models covered in this paper is of course not exhaustive and it would be interesting to know if our two-step approach could be deployed in other settings with "logit" noise. In ongoing work, we have found that this is one avenue to approach estimation of dynamic ordered logit models, potentially of arbitrary lag order.

References

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

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

Andrews, D. W. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666.

Arellano, M. and Carrasco, R. (2003). Binary choice panel data models with predetermined variables. *Journal of econometrics*, 115(1):125–157.

Arellano, M. and Honoré, B. (2001). Panel data models: some recent developments. In *Handbook of econometrics*, volume 5, pages 3229–3296. Elsevier.

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

Bonhomme, S., Dano, K., and Graham, B. S. (2023). Identification in a binary choice panel data model with a predetermined covariate. Technical report, National Bureau of Economic Research.

Browning, M. and Carro, J. M. (2014). Dynamic binary outcome models with maximal heterogeneity. *Journal of Econometrics*, 178(2):805–823.

Card, D. and Hyslop, D. R. (2005). Estimating the effects of a time-limited earnings subsidy for welfare-leavers. *Econometrica*, 73(6):1723–1770.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The review of economic studies*, 47(1):225–238.

Chamberlain, G. (1985). *Heterogeneity, omitted variable bias, and duration dependence*, page 3–38. Econometric Society Monographs. Cambridge University Press.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics*, 34(3):305–334.

Chay, K. Y., Hoynes, H. W., and Hyslop, D. (1999). A non-experimental analysis of true state dependence in monthly welfare participation sequences. In *American Statistical Association*, pages 9–17.

Chay, K. Y. and Hyslop, D. (1998). *Identification and estimation of dynamic binary response panel data models: empirical evidence using alternative approaches*. Number 5. Center for Labor Economics, University of California, Berkeley.

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

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Davezies, L., D'Haultfoeuille, X., and Laage, L. (2021). Identification and estimation of average marginal effects in fixed effects logit models. *arXiv preprint arXiv:2105.00879*.

Davezies, L., D'Haultfœuille, X., and Mugnier, M. (2023). Fixed-effects binary choice models with three or more periods. *Quantitative Economics*, 14(3):1105–1132.

Deza, M. (2015). Is there a stepping stone effect in drug use? separating state dependence from unobserved heterogeneity within and between illicit drugs. *Journal of Econometrics*, 184(1):193–207.

Dobronyi, C., Gu, J., et al. (2021). Identification of dynamic panel logit models with fixed effects. *arXiv preprint arXiv:2104.04590*.

Dobronyi, C. R., Ouyang, F., and Yang, T. T. (2023). Revisiting panel data discrete choice models with lagged dependent variables. *arXiv preprint arXiv:2301.09379*.

Dubé, J.-P., Hitsch, G. J., and Rossi, P. E. (2010). State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, 41(3):417–445.

Egger, P. H., Pfaffermayr, M., and Weber, A. (2003). Sectoral adjustment of employment: the impact of outsourcing and trade at the micro level. *Available at SSRN 469841*.

Graham, B. S. (2013). Comment on "social networks and the identification of peer effects" by paul goldsmith-pinkham and guido w. imbens. *Journal of Business and Economic Statistics*, 31(3):266–270.

Graham, B. S. (2016). Homophily and transitivity in dynamic network formation. Technical report, National Bureau of Economic Research.

Gu, J., Hahn, J., and Kim, K. I. (2023). The information bound of a dynamic panel logit model with fixed effects—corrigendum. *Econometric Theory*, 39(1):219–219.

Hahn, J. (1997). A note on the efficient semiparametric estimation of some exponential panel models. *Econometric Theory*, 13(4):583–588.

Hahn, J. (2001). The information bound of a dynamic panel logit model with fixed effects. *Econometric Theory*, 17(5):913–932.

Hansen, B. E. and Lee, S. (2021). Inference for iterated gmm under misspecification. *Econometrica*, 89(3):1419–1447.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Heckman, J. J. (1981). Heterogeneity and state dependence. In *Studies in labor markets*, pages 91–140. University of Chicago Press.

Honoré, B. E. and De Paula, Á. (2021). Identification in simple binary outcome panel data models. *The Econometrics Journal*, 24(2):C78–C93.

Honoré, B. E., Hu, L., Kyriazidou, E., and Weidner, M. (2022). Simultaneity in binary outcome models with an application to employment for couples. *arXiv preprint arXiv:2207.07343*.

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

Honoré, B. E. and Kyriazidou, E. (2019). Panel vector autoregressions with binary data. In *Panel Data Econometrics*, pages 197–223. Elsevier.

Honoré, B. E., Muris, C., and Weidner, M. (2021). Dynamic ordered panel logit models. *arXiv preprint arXiv:2107.03253*.

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

Honoré, B. E. and Weidner, M. (2020). Moment conditions for dynamic panel logit models with fixed effects. *arXiv preprint arXiv:2005.05942*.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*, 20(4):493–506.

Kitazawa, Y. (2022). Transformations and moment conditions for dynamic fixed effects logit models. *Journal of Econometrics*, 229(2):350 – 362.

Kitazawa, Y. et al. (2013). Exploration of dynamic fixed effects logit models from a traditional angle. Technical report.

Kitazawa, Y. et al. (2016). Root-n consistent estimations of time dummies for the dynamic fixed effects logit models: Monte carlo illustrations. Technical report.

Kruiniger, H. (2020). Further results on the estimation of dynamic panel logit models with fixed effects. *arXiv preprint arXiv:2010.03382*.

Magnac, T. (2000). Subsidised training and youth employment: distinguishing unobserved heterogeneity from state dependence in labour market histories. *The economic journal*, 110(466):805–837.

Muris, C., Raposo, P., and Vandoros, S. (2020). A dynamic ordered logit model with fixed effects. *arXiv preprint arXiv:2008.05517*.

Narendranthan, W., Nickell, S., and Metcalf, D. (1985). An investigation into the incidence and dynamic structure of sickness and unemployment in britain, 1965–75. *Journal of the Royal Statistical Society: Series A (General)*, 148(3):254–267.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135.

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

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

Pakel, C. and Weidner, M. (2023). Bounds on average effects in discrete choice panel data models. *arXiv preprint arXiv:2309.09299*.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, 90(1):77–97.

Appendix

A Partial Fraction Decomposition

Lemma 8. *For any reals $u_1, u_2, \dots, u_K$, $v_1, v_2, \dots, v_K$ and $a_1, a_2, \dots, a_K$, $K \geq 1$ we have*

$$\frac{1}{1+\sum\limits_{k=1}^{K} e^{v_k+a_k}} + \sum_{k=1}^{K}(1-e^{u_k-v_k})\frac{e^{v_k+a_k}}{\left(1+\sum\limits_{k=1}^{K} e^{v_k+a_k}\right)\left(1+\sum\limits_{k=1}^{K} e^{u_k+a_k}\right)} = \frac{1}{1+\sum\limits_{k=1}^{K} e^{u_k+a_k}}$$

and

$$\frac{e^{v_j+a_j}}{1+\sum\limits_{k=1}^{K} e^{v_k+a_k}} + (1-e^{-u_j+v_j})\frac{e^{u_j+a_j}}{\left(1+\sum\limits_{k=1}^{K} e^{v_k+a_k}\right)\left(1+\sum\limits_{k=1}^{K} e^{u_k+a_k}\right)} +$$
$$\sum_{\substack{k=1\\k\neq j}}^{K}(1-e^{(u_k-u_j)-(v_k-v_j)})\frac{e^{v_k+a_k+u_j+a_j}}{\left(1+\sum\limits_{k=1}^{K} e^{v_k+a_k}\right)\left(1+\sum\limits_{k=1}^{K} e^{u_k+a_k}\right)} = \frac{e^{u_j+a_j}}{1+\sum\limits_{k=1}^{K} e^{u_k+a_k}}$$

Proof.

$$\begin{aligned}
&\frac{1}{1+\sum_{k=1}^{K} e^{v_k+a_k}} + \sum_{k=1}^{K}(1-e^{u_k-v_k})\frac{e^{v_k+a_k}}{\left(1+\sum_{k=1}^{K} e^{v_k+a_k}\right)\left(1+\sum_{k=1}^{K} e^{u_k+a_k}\right)} \\
&= \frac{1+\sum_{k=1}^{K} e^{u_k+a_k} + \sum_{k=1}^{K} e^{v_k+a_k} - \sum_{k=1}^{K} e^{u_k+a_k}}{\left(1+\sum_{k=1}^{K} e^{v_k+a_k}\right)\left(1+\sum_{k=1}^{K} e^{u_k+a_k}\right)} \\
&= \frac{1+\sum_{k=1}^{K} e^{v_k+a_k}}{\left(1+\sum_{k=1}^{K} e^{v_k+a_k}\right)\left(1+\sum_{k=1}^{K} e^{u_k+a_k}\right)} \\
&= \frac{1}{1+\sum_{k=1}^{K} e^{u_k+a_k}}
\end{aligned}$$

and

$$
\begin{aligned}
&\frac{e^{v_j+a_j}}{1+\sum_{k=1}^K e^{v_k+a_k}} + (1-e^{-u_j+v_j})\frac{e^{u_j+a_j}}{\left(1+\sum_{k=1}^K e^{v_k+a_k}\right)\left(1+\sum_{k=1}^K e^{u_k+a_k}\right)} + \\
&\sum_{\substack{k=1\\k\neq j}}^K (1-e^{(u_k-u_j)-(v_k-v_j)})\frac{e^{v_k+a_k+u_j+a_j}}{\left(1+\sum_{k=1}^K e^{v_k+a_k}\right)\left(1+\sum_{k=1}^K e^{u_k+a_k}\right)} \\
&= \frac{e^{v_j+a_j} + \sum_{k=1}^K e^{v_j+a_j+u_k+a_k} + e^{u_j+a_j} - e^{v_j+a_j} + \sum\limits_{\substack{k=1\\k\neq j}}^K e^{v_k+a_k+u_j+a_j} - \sum_{\substack{k=1\\k\neq j}}^K e^{v_j+a_j+u_k+a_k}}{\left(1+\sum_{k=1}^K e^{v_k+a_k}\right)\left(1+\sum_{k=1}^K e^{u_k+a_k}\right)} \\
&= \frac{e^{u_j+a_j} + e^{v_j+a_j+u_j+a_j} + \sum\limits_{\substack{k=1\\k\neq j}}^K e^{v_k+a_k+u_j+a_j}}{\left(1+\sum_{k=1}^K e^{v_k+a_k}\right)\left(1+\sum_{k=1}^K e^{u_k+a_k}\right)} \\
&= \frac{e^{u_j+a_j}\left(1+\sum\limits_{k=1}^K e^{v_k+a_k}\right)}{\left(1+\sum_{k=1}^K e^{v_k+a_k}\right)\left(1+\sum_{k=1}^K e^{u_k+a_k}\right)} \\
&= \frac{e^{u_j+a_j}}{1+\sum_{k=1}^K e^{u_k+a_k}}
\end{aligned}
$$

□

Lemma 9. *Fix $M \geq 2$, let $\mathcal{Y} = \{0,1\}^M$. Then, for any $k \in \mathcal{Y}$ and any reals $u_1, u_2, \ldots, u_M$, $v_1, v_2, \ldots, v_M$ and $a_1, a_2, \ldots, a_M$, we have*

$$
\prod_{m=1}^M \frac{e^{k_m(v_m+a_m)}}{1+e^{v_m+a_m}} + \sum_{l\in\mathcal{Y}\setminus\{k\}} \left[1-e^{\sum_{j=1}^M (l_j-k_j)(u_j-v_j)}\right] \prod_{m=1}^M \frac{e^{k_m(u_m+a_m)}}{1+e^{u_m+a_m}} \frac{e^{l_m(v_m+a_m)}}{1+e^{v_m+a_m}} = \prod_{m=1}^M \frac{e^{k_m(u_m+a_m)}}{1+e^{u_m+a_m}}
$$

Proof. Let

$$
LHS = \prod_{m=1}^M \frac{e^{k_m(v_m+a_m)}}{1+e^{v_m+a_m}} + \sum_{l\in\mathcal{Y}\setminus\{k\}} \left[1-e^{\sum_{j=1}^M (l_j-k_j)(u_j-v_j)}\right] \prod_{m=1}^M \frac{e^{k_m(u_m+a_m)}}{1+e^{u_m+a_m}} \frac{e^{l_m(v_m+a_m)}}{1+e^{v_m+a_m}}
$$

and let Num denote the numerator of LHS. We have:

$$\begin{aligned}
Num &= Num_1 + Num_2 \\
Num_1 &= \prod_{m=1}^{M} e^{k_m(v_m+a_m)}(1+e^{u_m+a_m}) \\
Num_2 &= \sum_{l\in\mathcal{Y}\setminus\{k\}} \left[1 - e^{\sum_{j=1}^{M}(l_j-k_j)(u_j-v_j)}\right] \prod_{m=1}^{M} e^{k_m(u_m+a_m)+l_m(v_m+a_m)} \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)} \sum_{l\in\mathcal{Y}\setminus\{k\}} \prod_{m=1}^{M} e^{l_m(v_m+a_m)} - \sum_{l\in\mathcal{Y}\setminus\{k\}} e^{\sum_{j=1}^{M} l_j(u_j+a_j)+k_j(v_j+a_j)} \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)} \sum_{l\in\mathcal{Y}\setminus\{k\}} \prod_{m=1}^{M} e^{l_m(v_m+a_m)} - \prod_{m=1}^{M} e^{k_m(v_m+a_m)} \sum_{l\in\mathcal{Y}\setminus\{k\}} \prod_{m=1}^{M} e^{l_m(u_m+a_m)}
\end{aligned}$$

Now, noting that

$$\begin{aligned}
\sum_{l\in\mathcal{Y}} \prod_{m=1}^{M} e^{l_m(v_m+a_m)} &= \prod_{m=1}^{M}(1+e^{v_m+a_m}) \\
\sum_{l\in\mathcal{Y}} \prod_{m=1}^{M} e^{l_m(u_m+a_m)} &= \prod_{m=1}^{M}(1+e^{u_m+a_m})
\end{aligned}$$

we get

$$\begin{aligned}
Num_2 &= \prod_{m=1}^{M} e^{k_m(u_m+a_m)} \sum_{l\in\mathcal{Y}\setminus\{k\}} \prod_{m=1}^{M} e^{l_m(v_m+a_m)} - \prod_{m=1}^{M} e^{k_m(v_m+a_m)} \sum_{l\in\mathcal{Y}\setminus\{k\}} \prod_{m=1}^{M} e^{l_m(u_m+a_m)} \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)} \left(\prod_{m=1}^{M}(1+e^{v_m+a_m}) - \prod_{m=1}^{M} e^{k_m(v_m+a_m)}\right) \\
&- \prod_{m=1}^{M} e^{k_m(v_m+a_m)} \left(\prod_{m=1}^{M}(1+e^{u_m+a_m}) - \prod_{m=1}^{M} e^{k_m(u_m+a_m)}\right) \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)}(1+e^{v_m+a_m}) - \prod_{m=1}^{M} e^{k_m(v_m+a_m)}(1+e^{u_m+a_m}) \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)}(1+e^{v_m+a_m}) - Num_1
\end{aligned}$$

It follows that $Num = \prod_{m=1}^{M} e^{k_m(u_m+a_m)}(1+e^{v_m+a_m})$ and consequently

$$LHS = \frac{\prod_{m=1}^{M} e^{k_m(u_m+a_m)}(1+e^{v_m+a_m})}{\prod_{m=1}^{M}(1+e^{u_m+a_m})(1+e^{v_m+a_m})} = \prod_{m=1}^{M} \frac{e^{k_m(u_m+a_m)}}{1+e^{u_m+a_m}}$$

$\square$

B Connection to Kitazawa and Honoré-Weidner

Recall from Proposition 2 that when $T \geq 3$, our simplest moment conditions for t, s such that $T-1 \geq t > s \geq 1$ write:

$$\begin{aligned}
\psi_\theta^{0|0}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) &= \phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) - \zeta_\theta^{0|0}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) \\
&= \phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) - (1-Y_{is}) - \omega_{t,s}^{0|0}(\theta) Y_{is} \phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) \\
\psi_\theta^{1|1}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) &= \phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) - \zeta_\theta^{1|1}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) \\
&= \phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) - Y_{is} - \omega_{t,s}^{1|1}(\theta)(1-Y_{is}) \phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)
\end{aligned}$$

where we know from Lemma 3 that

$$\begin{aligned}
\omega_{t,s}^{0|0}(\theta) &= 1 - e^{(\kappa_t^{0|0}(\theta) - \mu_s(\theta))} \\
&= 1 - e^{(X_{it+1} - X_{is})'\beta - \gamma Y_{is-1}} \\
\omega_{t,s}^{1|1}(\theta) &= 1 - e^{-(\kappa_t^{1|1}(\theta) - \mu_s(\theta))} \\
&= 1 - e^{-\gamma(1-Y_{is-1}) - (X_{it+1} - X_{is})'\beta)}
\end{aligned}$$

Now, note that:

$$\begin{aligned}
\tanh\left(\frac{\gamma(1-Y_{it-2}) + (\Delta X_{it} + \Delta X_{it+1})'\beta}{2}\right) &= \frac{1 - e^{-\left(\gamma(1-Y_{it-2}) + (\Delta X_{it} + \Delta X_{it+1})'\beta\right)}}{1 + e^{-\left(\gamma(1-Y_{it-2}) + (\Delta X_{it} + \Delta X_{it+1})'\beta\right)}} = \frac{\omega_{t,t-1}^{1|1}(\theta)}{2 - \omega_{t,t-1}^{1|1}(\theta)} \\
\tanh\left(\frac{-\gamma Y_{it-2} + (\Delta X_{it} + \Delta X_{it+1})'\beta}{2}\right) &= \frac{e^{-\gamma Y_{it-2} + (\Delta X_{it} + \Delta X_{it+1})'\beta} - 1}{e^{-\gamma Y_{it-2} + (\Delta X_{it} + \Delta X_{it+1})'\beta} + 1} = -\frac{\omega_{t,t-1}^{0|0}(\theta)}{2 - \omega_{t,t-1}^{0|0}(\theta)}
\end{aligned}$$

and $\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = \Upsilon_{it}$ and $1 - \phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = U_{it}$. Thus, we have:

$$
\begin{aligned}
(2-\omega^{0|0}_{t,t-1}(\theta))\hbar U_{it} &= (2-\omega^{0|0}_{t,t-1}(\theta))(U_{it}-Y_{it-1}) + \omega^{0|0}_{t,t-1}(\theta)\left(U_{it}+Y_{it-1}-2U_{it}Y_{it-1}\right) \\
&= 2\left[U_{it}-Y_{it-1}+\omega^{0|0}_{t,t-1}(\theta)Y_{it-1}(1-U_{it})\right] \\
&= 2\left[1-\phi^{0|0}_{\theta}(Y_{it+1},Y_{it},Y_{it-1},X_i)-Y_{it-1}+\omega^{0|0}_{t,t-1}(\theta)Y_{it-1}\phi^{0|0}_{\theta}(Y_{it+1},Y_{it},Y_{it-1},X_i)\right] \\
&= -2\left[\phi^{0|0}_{\theta}(Y_{it+1},Y_{it},Y_{it-1},X_i)-(1-Y_{it-1})-\omega^{0|0}_{t,t-1}(\theta)Y_{it-1}\phi^{0|0}_{\theta}(Y_{it+1},Y_{it},Y_{it-1},X_i)\right] \\
&= -2\psi^{0|0}_{\theta}(Y^{t+1}_{it-1},Y^{t-1}_{it-2},X_i) \\
(2-\omega^{1|1}_{t,t-1}(\theta))\hbar \Upsilon_{it} &= (2-\omega^{1|1}_{t,t-1}(\theta))(\Upsilon_{it}-Y_{it-1}) - \omega^{1|1}_{t,t-1}(\theta)\left(\Upsilon_{it}+Y_{it-1}-2\Upsilon_{it}Y_{it-1}\right) \\
&= 2\left[\Upsilon_{it}-Y_{it-1}-\omega^{1|1}_{t,t-1}(\theta)\Upsilon_{it}\left(1-Y_{it-1}\right)\right] \\
&= 2\left[\phi^{1|1}_{\theta}(Y_{it+1},Y_{it},Y_{it-1},X_i)-Y_{it-1}-\omega^{1|1}_{t,t-1}(\theta)\phi^{1|1}_{\theta}(Y_{it+1},Y_{it},Y_{it-1},X_i)\left(1-Y_{it-1}\right)\right] \\
&= 2\psi^{1|1}_{\theta}(Y^{t+1}_{it-1},Y^{t-1}_{it-2},X_i)
\end{aligned}
$$

To establish the connection to the work of Honoré and Weidner (2020), it is useful to re-write the moment functions slightly differently. By re-arranging terms, one obtains the following for $T=3$

$$
\begin{aligned}
\psi^{0|0}_{\theta}(Y^3_1,Y^1_{i0},X_i) &= (1-Y_{i1})\phi^{0|0}_{\theta}(Y^3_{i1},X_i) + e^{(X_{i3}-X_{i1})'\beta-\gamma Y_{i0}}Y_{i1}\phi^{0|0}_{\theta}(Y^3_{i1},X_i) - (1-Y_{i1}) \\
&= e^{(X_{i2}-X_{i3})'\beta}(1-Y_{i1})(1-Y_{i2})Y_{i3} + (1-Y_{i1})(1-Y_{i2})(1-Y_{i3}) \\
&+ e^{(X_{i2}-X_{i1})'\beta+\gamma(1-Y_{i0})}Y_{i1}(1-Y_{i2})Y_{i3} \\
&+ e^{(X_{i3}-X_{i1})'\beta-\gamma Y_{i0}}Y_{i1}(1-Y_{i2})(1-Y_{i3}) \\
&- (1-Y_{i1}) \qquad (9) \\
&= (e^{(X_{i2}-X_{i3})'\beta}-1)(1-Y_{i1})(1-Y_{i2})Y_{i3} \\
&+ e^{(X_{i2}-X_{i1})'\beta+\gamma(1-Y_{i0})}Y_{i1}(1-Y_{i2})Y_{i3} \\
&+ e^{(X_{i3}-X_{i1})'\beta-\gamma Y_{i0}}Y_{i1}(1-Y_{i2})(1-Y_{i3}) \\
&- (1-Y_{i1})Y_{i2}
\end{aligned}
$$

where the last line uses the fact that: $(1-Y_{i1}) = (1-Y_{i1})Y_{i2} + (1-Y_{i1})(1-Y_{i2})Y_{i3} + (1-Y_{i1})(1-Y_{i2})(1-Y_{i3})$ to make some cancellations. For the initial condition, $Y_{i0}=0$, equation (9) corresponds to their moment function m^b_0 which they express in an extensive form. For

$Y_{i0} = 1$, we get instead m_1^b. Similarly,

$$\begin{aligned}
\psi_\theta^{1|1}(Y_{i1}^3, Y_{i0}^1, X_i) &= Y_{i1}\phi_\theta^{1|1}(Y_{i1}^3, X_i) + e^{-\gamma(1-Y_{i0})-(X_{i3}-X_{i1})'\beta)}(1-Y_{i1})\phi_\theta^{1|1}(Y_{i1}^3, X_i) - Y_{i1} \\
&= e^{(X_{i3}-X_{i2})'\beta}Y_{i1}Y_{i2}(1-Y_{i3}) + Y_{i1}Y_{i2}Y_{i3} \\
&+ e^{(X_{i1}-X_{i2})'\beta+\gamma Y_{i0}}(1-Y_{i1})Y_{i2}(1-Y_{i3}) \\
&+ e^{(X_{i1}-X_{i3})'\beta-\gamma(1-Y_{i0})}(1-Y_{i1})Y_{i2}Y_{i3} \\
&- Y_{i1} \qquad (10)\\
&= (e^{(X_{i3}-X_{i2})'\beta} - 1)Y_{i1}Y_{i2}(1-Y_{i3}) \\
&+ e^{(X_{i1}-X_{i2})'\beta+\gamma Y_{i0}}(1-Y_{i1})Y_{i2}(1-Y_{i3}) \\
&+ e^{(X_{i1}-X_{i3})'\beta-\gamma(1-Y_{i0})}(1-Y_{i1})Y_{i2}Y_{i3} \\
&- Y_{i1}(1-Y_{i2})
\end{aligned}$$

where the last line uses the fact that: $Y_{i1} = Y_{i1}(1-Y_{i2}) + Y_{i1}Y_{i2}Y_{i3} + Y_{i1}Y_{i2}(1-Y_{i3})$. For the initial condition $Y_{i0} = 0$, equation (10) gives their moment function m_0^a and for $Y_{i0} = 1$, we get m_1^a. Our moments are thus identical, at least for the case $T = 3$.

C The remaining steps for the AR(p) model with $p > 1$

As indicated in Section 4.4.2 , **Step 1)** (b) is now analogous to the AR(1) case since the transition probabilities keep an identical structure. As soon as $T \geq p+2$, we can construct transition functions other than $\phi_\theta^{y_1|y_1^p}(Y_{it+1}, Y_{it}, Y_{it-(2p-1)}^{t-1}, X_i)$ also associated to $\pi_t^{y_1|y_1^p}(A_i, X_i)$, for $y_1^p \in \mathcal{Y}^p$ in periods $t \in \{p+1, \ldots, T-1\}$,. These new transition functions that we denote $\zeta_\theta^{y_1|y_1^p}(.)$ take the form of a weighted combination of past outcome $\mathbb{1}(Y_{is} = y_1)$, $s \in \{1, \ldots, t-p\}$ and the interaction of $\mathbb{1}(Y_{is} \neq y_1)$ with any transition function whose conditioning set encompasses Y_{is} for it to map to $\pi_t^{y_1|y_1^p}(A_i, X_i)$. The simplest examples which are also the only ones available when $T = p+2$, are given in Lemma 10.

Lemma 10. *In model (5) with $T \geq p+2$, for all $t \in \{p+1, \ldots, T-1\}$, $s \in \{1, \ldots, t-p\}$*

and $y_1^p \in \mathcal{Y}^p$*, let*

$$\mu_s(\theta) = \sum_{r=1}^{p} \gamma_{0r} Y_{is-r} + X_{is}'\beta$$
$$\kappa_t^{y_1|y_1^p}(\theta) = \sum_{r=1}^{p} \gamma_{0r} y_r + X_{it+1}'\beta$$
$$\omega_{t,s}^{y_1|y_1^p}(\theta) = \left[1 - e^{(\kappa_t^{y_1|y_1^p}(\theta)-\mu_s(\theta))}\right]^{1-y_1} \left[1 - e^{-(\kappa_t^{y_1|y_1^p}(\theta)-\mu_s(\theta))}\right]^{y_1}$$

and define the moment functions:

$$\zeta_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is-p}^{s}, X_i) = \mathbb{1}\{Y_{is} = y_1\} + \omega_{t,s}^{y_1|y_1^p}(\theta)\mathbb{1}\{Y_{is} \neq y_1\}\phi_\theta^{y_1|y_1^p}(Y_{it+1}, Y_{it}, Y_{it-(2p-1)}^{t-1}, X_i)$$

Then,

$$\mathbb{E}\left[\zeta_{\theta_0}^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is-p}^{s}, X_i)|Y_i^0, Y_{i1}^{s-1}, X_i, A_i\right] = \pi_t^{y_1|y_1^p}(A_i, X_i)$$

Unsurprisingly, as in the AR(1) case, it becomes possible to construct iteratively more transition functions from those given in Lemma 10 when at least $T = p+3$ periods are observed post initial condition. They are given in Corollary 10.1 below.

Corollary 10.1. *In model (5) with* $T \geq p+3$*, for all* $t \in \{p+1, \ldots, T-1\}$ *and collection of ordered indices* s_1^J *with* $J \geq 2$ *satisfying* $t-p \geq s_1 > \ldots > s_J \geq 1$*, and for all* $y_1^p \in \mathcal{Y}^p$*, let*

$$\zeta_\theta^{0|0,y_2^p}(Y_{it-(2p-1)}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_J-p}^{s_J}, X_i) = (1 - Y_{is_J}) + \omega_{t,s_J}^{0|0,y_2^p}(\theta) Y_{is_J} \zeta_\theta^{0|0,y_2^p}(Y_{it-1}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_{J-1}-p}^{s_{J-1}}, X_i)$$
$$\zeta_\theta^{1|1,y_2^p}(Y_{it-(2p-1)}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_J-p}^{s_J}, X_i) = Y_{is_j} + \omega_{t,s_J}^{1|1,y_2^p}(\theta)(1 - Y_{is_J}) \zeta_\theta^{1|1,y_2^p}(Y_{it-1}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_{J-1}-p}^{s_{J-1}}, X_i)$$

with weights $\omega_{t,s_J}^{y_1|y_1^p}(\theta)$ *defined as in Lemma 10. Then,*

$$\mathbb{E}\left[\zeta_{\theta_0}^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_J-p}^{s_J}, X_i)|Y_i^0, Y_{i1}^{s_J-1}, X_i, A_i\right] = \pi_t^{y_1|y_1^p}(A_i, X_i)$$

Step 2). Provided that $T \geq p+2$, it is clear that the difference between any two distinct transition functions associated to the same transition probability in $t \in \{p+1, \ldots, T-1\}$ will yield a valid moment function. Proposition 4 hereinbelow presents one set of valid moment functions that generalize those obtained previously for the one lag case.

Proposition 4. *In model (5)*

if $T \geq p+2$, for all $t \in \{p+1, \ldots, T-1\}$, $s \in \{1, \ldots, t-p\}$ and $y_1^p \in \mathcal{Y}^p$, let

$$\psi_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is-p}^s, X_i) = \phi_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, X_i) - \zeta_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is-p}^s, X_i),$$

if $T \geq p+3$, for all $t \in \{p+1, \ldots, T-1\}$ and collection of ordered indices s_1^J with $J \geq 2$ satisfying $t-p \geq s_1 > \ldots > s_J \geq 1$, and for all $y_1^p \in \mathcal{Y}^p$, let

$$\psi_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_J-p}^{s_J}, X_i) = \phi_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, X_i) - \zeta_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_J-p}^{s_J}, X_i)$$

Then,

$$\mathbb{E}\left[\psi_{\theta_0}^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is-p}^s, X_i)|Y_i^0, Y_{i1}^{s-1}, X_i, A_i\right] = 0$$
$$\mathbb{E}\left[\psi_{\theta_0}^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_J-p}^{s_J}, X_i)|Y_i^0, Y_{i1}^{s_J-1}, X_i, A_i\right] = 0$$

This family of moment functions features precisely $2^T - (T+1-p)2^p$ distinct elements for any initial condition. Indeed, fix Y_i^0 and a p-vector $y_1^p \in \{0,1\}^p$. Then, for a given time period $t \in \{p+1, \ldots, T-1\}$, there are $\binom{t-p}{1}$ moments of the form $\psi_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is-p}^s, X_i)$ corresponding to choices of $s \in \{1, \ldots, t-p\}$. Moreover, by choosing any feasible sequence s_1^J, $J \geq 2$, verifying $t-p \geq s_1 > \ldots > s_J \geq 1$ we produce another $\sum_{l=2}^{t-p}\binom{t-p}{l}$ moment functions of the form $\psi_\theta^{y_1|y_1^p}(Y_{it-(2p-1)}^{t+1}, Y_{is_1-p}^{s_1}, \ldots, Y_{is_J-p}^{s_J}, X_i)$. In total, for period t, we count :

$$\sum_{l=1}^{t-p}\binom{t-p}{l} = 2^{t-p} - 1$$

valid moments. Now, summing over all possible values for $t \in \{p+1, \ldots, T-1\}$ and multiplying by the number of distinct values for y_1^p, namely 2^p, we get:

$$2^p \sum_{t=p+1}^{T-1}\sum_{l=1}^{t-p}\binom{t-p}{l} = 2^p \sum_{t=p+1}^{T-1}(2^{t-p}-1) = 2^p\left(2\frac{1-2^{T-p-1}}{1-2} - (T-p-1)\right) = 2^T - (T+1-p)2^p$$

Numerical experimentation for various values of T in the AR(1) and AR(2) cases suggest that the moment functions of Proposition 4 are effectively linearly independent. Therefore, Theorem 3 implies that they constitute a complete family of moment functions for AR(p) models. From a practical standpoint, this shows that functional differencing at least in panel data logit models can be broken down into a series of equivalent simpler subproblems period by period that find all moment equality restrictions. Our procedure can be advantageous in sophisticated models with a few lags where an analysis of the full likelihood, a high dimensional object, can prove difficult.

D Simulation Experiments

In this section, we report the results of a small set of simulations designed to assess the finite sample performance of GMM estimators based on our moment conditions.

D.1 Monte Carlo for an AR(3) logit model

For our first example, we consider an AR(3) logit model with $T = 5$ periods (i.e 8 periods in total with the initial condition) and a single exogenous covariate. We set the common parameters to $\gamma_{01} = 1.0$, $\gamma_{02} = 0.5, \gamma_{03} = 0.25, \beta_0 = 0.5$ and use the following generative model in the spirit of Honoré and Kyriazidou (2000):

$$
\begin{aligned}
Y_{i-2} &= \mathbb{1}\{X_{i-2}'\beta_0 + A_i - \epsilon_{i-2} \geq 0\} \\
Y_{i-1} &= \mathbb{1}\{\gamma_{01}Y_{i-2} + X_{i-1}'\beta_0 + A_i - \epsilon_{i-1} \geq 0\} \\
Y_{i0} &= \mathbb{1}\{\gamma_{01}Y_{i-1} + \gamma_{02}Y_{i-2} + X_{i0}'\beta_0 + A_i - \epsilon_{i0} \geq 0\} \\
Y_{it} &= \mathbb{1}\left\{\gamma_{01}Y_{it-1} + \gamma_{02}Y_{it-2} + \gamma_{03}Y_{it-3} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\right\}, \quad t = 1, \ldots, 5
\end{aligned}
$$

The disturbances ϵ_{it} are iid standard logistic over time, X_{it} is iid $\mathcal{N}(0,1)$ and the fixed effects are computed as $A_i = \frac{1}{\sqrt{8}} \sum_{t=-2}^{5} X_{it}$. To evaluate the performance of the estimators described below, we simulate data for four sample sizes : 500, 2000, 8000, 16000, and perform 1000 Monte Carlo replications for each design.

For $T = 5$, we know from Proposition 4 that 8 valid moment functions are available, each stemming from the 8 possible transition probabilities of the model (there are really 16 transition probabilities in total but 8 are redundant since probabilities sum to one). We consider the interaction of all 8 valid moment functions with a constant, the 3 initial conditions Y_{i-2}, Y_{i-1}, Y_{i0} and the covariates X_{it} in each period $t \in \{1, \ldots, 5\}$ to construct the

72×1 moment vector:

$$m_\theta(Y_i, Y_i^0, X_i) = \begin{pmatrix} \psi_\theta^{0|0,0,0}(Y_{i-1}^5, Y_{i-2}^1, X_i) \\ \psi_\theta^{0|0,0,1}(Y_{i-1}^5, Y_{i-2}^1, X_i) \\ \psi_\theta^{0|0,1,0}(Y_{i-1}^5, Y_{i-2}^1, X_i) \\ \psi_\theta^{0|0,1,1}(Y_{i-1}^5, Y_{i-2}^1, X_i) \\ \psi_\theta^{1|1,0,0}(Y_{i-1}^5, Y_{i-2}^1, X_i) \\ \psi_\theta^{1|1,0,1}(Y_{i-1}^5, Y_{i-2}^1, X_i) \\ \psi_\theta^{1|1,1,0}(Y_{i-1}^5, Y_{i-2}^1, X_i) \\ \psi_\theta^{1|1,1,1}(Y_{i-1}^5, Y_{i-2}^1, X_i) \end{pmatrix} \otimes \begin{pmatrix} 1 \\ Y_{i-2} \\ Y_{i-1} \\ Y_{i0} \\ X_{i1}^{5'} \end{pmatrix}$$

where $\otimes$ denotes the standard Kronecker product. The choice of this particular set of instruments is of course arbitrary and only motivated by simplicity. We also consider a rescaled version of $m_\theta(Y_i, Y_i^0, X_i)$ that we denote $\widetilde{m_\theta}(Y_i, Y_i^0, X_i)$ where each of the 8 valid moment functions are appropriately rescaled so that $\forall y_1^3 \in \{0,1\}^3$, $\sup_{X_i, Y_i, \theta} \left| \psi_\theta^{y_1|y_1,y_2,y_3}(Y_{i-1}^5, Y_{i-2}^1, X_i) \right| < \infty$. We do so by normalizing $\psi_\theta^{y_1|y_1,y_2,y_3}(Y_{i-1}^5, Y_{i-2}^1, X_i)$ by the sum of the absolute values of all unique values it can take as a function over choice histories Y_{i1}^5. The rationale for normalizing the moments originates from Honoré and Weidner (2020) who presented numerical evidence that a rescaling of this kind improved the finite sample performance of their estimators in the one and two lags cases. Given, $m_\theta(Y_i, Y_i^0, X_i)$ and $\widetilde{m_\theta}(Y_i, Y_i^0, X_i)$, we study the properties of two simple GMM estimators:

$$\hat{\theta}^a = \arg\max_{\theta \in \mathbb{R}^4} \left(\frac{1}{N} \sum_{i=1}^N m_\theta(Y_i, Y_i^0, X_i) \right)' \left(\frac{1}{N} \sum_{i=1}^N m_\theta(Y_i, Y_i^0, X_i) \right)$$

$$\hat{\theta}^b = \arg\max_{\theta \in \mathbb{R}^4} \left(\frac{1}{N} \sum_{i=1}^N \widetilde{m_\theta}(Y_i, Y_i^0, X_i) \right)' \left(\frac{1}{N} \sum_{i=1}^N \widetilde{m_\theta}(Y_i, Y_i^0, X_i) \right)$$

which both put equal weight on their individual components (i.e the weight matrix is the identity)[19]. Under standard regularity conditions, $\hat{\theta}^a, \hat{\theta}^b$ should be consistent and asymptotically normal.

[19]In a previous version of this paper we also considered a two-step "rescaled" estimator that uses a diagonal weight matrix with the inverse variance of each component in the spirit of Honoré and Weidner (2020). It performs very similarly to the equally-weighted estimator $\hat{\theta}^b$.

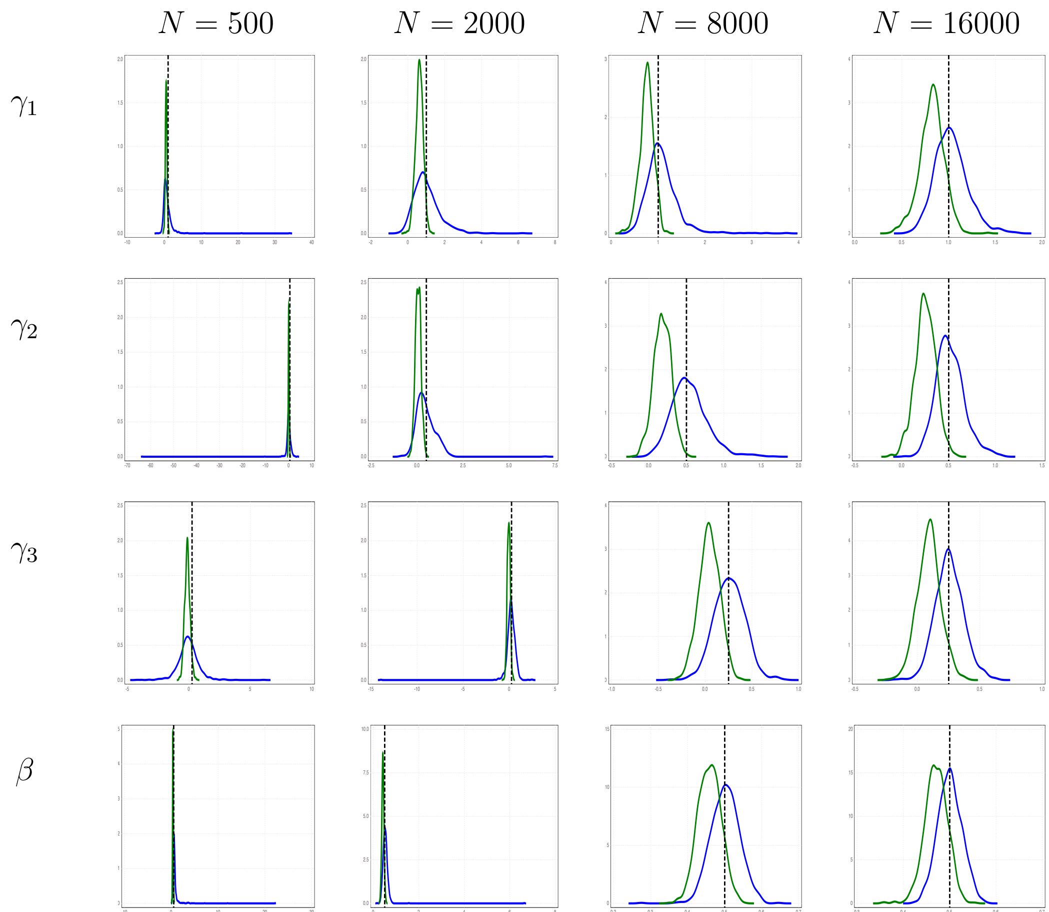
Table 2: Performance of GMM estimators for the AR(3)

		$\hat{\gamma}_1{}^a$	$\hat{\gamma}_1{}^b$	$\hat{\gamma}_2{}^a$	$\hat{\gamma}_2{}^b$	$\hat{\gamma}_3{}^a$	$\hat{\gamma}_3{}^b$	$\hat{\beta}^a$	$\hat{\beta}^b$
$N = 500$									
	Bias	-0.52	-0.50	-0.51	-0.50	-0.39	-0.32	-0.15	0.10
	MAE	0.52	0.69	0.51	0.58	0.39	0.51	0.15	0.14
$N = 2000$									
	Bias	-0.37	-0.10	-0.45	-0.12	-0.31	-0.04	-0.08	0.02
	MAE	0.37	0.42	0.45	0.34	0.31	0.25	0.08	0.06
$N = 8000$									
	Bias	-0.24	0.04	-0.32	0.01	-0.21	0.01	-0.04	0.00
	MAE	0.24	0.17	0.32	0.15	0.21	0.11	0.04	0.03
$N = 16000$									
	Bias	-0.18	0.01	-0.25	0.00	-0.16	0.00	-0.03	0.00
	MAE	0.18	0.11	0.25	0.10	0.16	0.07	0.03	0.02

NOTES: *Bias and MAE stand for median bias and median absolute error respectively. Reported results are based on a 1000 replications of the DGP.*

Table 2 presents the median bias and median absolute errors of the two GMM estimators for each design $N \in \{500, 2000, 8000, 16000\}$. Figure 1 plots their densities which as expected resemble gaussian distributions for the larger values of N. Interestingly, a first observation is that both estimators appear to suffer from a negative bias on the lag parameters at least up to $N = 2000$. And while this bias effectively vanishes for the "rescaled" GMM estimators for the larger sample size $N \geq 8000$, it remains quite significant for all lag parameters and also the slope coefficient for the "unnormalized" estimator. This is evident from the sign of the bias in Table 2 and from the fact that all green densities are to the left of the true parameters in Figure 1. This observation confirms the practical importance of normalizing all valid moment functions in binary response logit models to obtain precise estimates in small samples. Focusing on the "rescaled" estimator $\hat{\theta}^b$, we can see that it performs relatively well for $N \geq 8000$ with very little bias. This is corroborated in Figure 1: the blue densities are approximately centered at the true parameter values for $N \geq 8000$. Estimates for the slope parameter β are quite accurate even for $N = 500$ but precise estimation of the transition parameters requires a larger sample size. In terms of median absolute bias, it is interesting to note a ranking on the precision of estimates of the transition parameters: the coefficient on the first lag is noisier than the coefficient on the second lag which itself is noisier than

Figure 1: Densities of GMM estimators for the AR(3) with one regressor



Notes: The densities of estimates based on the first GMM estimator (i.e $\hat{\theta}^a$), the second GMM estimator (i.e $\hat{\theta}^b$) are indicated in green and blue respectively. Reported results are based on a 1000 replications of the DGP presented above with $\gamma_{01} = 1.0, \gamma_{02} = 0.5, \gamma_{03} = 0.25, \beta_0 = 0.5$. True parameter values are indicated with a vertical dashed line.

the coefficient on the third lag for each $N \in \{500, 2000, 8000, 16000\}$. In an unreported set of simulations, we have found that this empirical pattern is robust to other choices of the population parameters and initial condition and also applies to the AR(2) model with a similar data generating process.

D.2 Monte Carlo for a VAR(1) logit model

In our next example, we examine a bivariate VAR(1) logit model with $T = 3$ and scalar regressors $X_{m,it}$ in each layer $m \in \{1, 2\}$. We set the common parameters to $\gamma_{011} = \gamma_{022} = 1.0$, $\gamma_{012} = \gamma_{021} = 0.5$, $\beta_1 = \beta_2 = 0.5$. The data generating process is:

$$
\begin{aligned}
Y_{m,i0} &= \mathbb{1}\left\{X'_{m,i0}\beta_{0m} + A_{m,i} - \epsilon_{m,it} \geq 0\right\}, \quad m = 1,2 \\
Y_{m,it} &= \mathbb{1}\left\{\gamma_{0m1}Y_{1,it-1} + \gamma_{0m2}Y_{2,it-1} + X'_{m,it}\beta_{0m} + A_{m,i} - \epsilon_{m,it} \geq 0\right\}, \quad m = 1,2, \quad t = 1,2,3
\end{aligned}
$$

where the disturbances $\epsilon_{m,it}$ are iid standard logistic, the covariates $X_{m,it}$ are iid $\mathcal{N}(0,1)$ and the fixed effects are computed as $A_{m,i} = \frac{1}{\sqrt{4}}\sum_{t=0}^{3} X_{m,it}$. We consider sample sizes $N \in \{2000, 8000, 16000\}$ with 1000 Monte Carlo replications per design.

We use all four valid moment functions implied by Proposition 2 when $T = 3$ for the VAR(1) case, viz $\psi_\theta^{k|k}(Y_{i1}^3, Y_{i0}^1, X_i), k \in \{(0,0),(0,1),(1,0),(0,0)\}$ and form the 40×1 moment vector:

$$
m_\theta(Y_i, Y_i^0, X_i) = \begin{pmatrix} \psi_\theta^{(0,0)|(0,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(0,1)|(0,1)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,0)|(1,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,1)|(1,1)}(Y_{i1}^3, Y_{i0}^1, X_i) \end{pmatrix} \otimes \begin{pmatrix} 1 \\ Y_i^{0\prime} \\ X_{1,i1}^{3\prime} \\ X_{2,i1}^{3\prime} \end{pmatrix}
$$

Given the importance of rescaling the valid moment functions for better precision of GMM in the context of the AR(3), we also consider a normalized moment vector $\widetilde{m_\theta}(Y_i, Y_i^0, X_i)$ in which each $\psi_\theta^{k|k}(Y_{i1}^3, Y_{i0}^1, X_i)$ is divided by the sum of the absolute values of their unique non-zero entries as a 64-dimensional vector (64 possible choice histories Y_{i1}^3 per initial condition). With these moment functions in hand, we then compare the finite sample properties of three estimators: i) the VAR(1) analogs of $\hat{\theta}^a$ and $\hat{\theta}^b$ defined previously for the AR(3), ii) the iterated GMM estimator $\hat{\theta}^c$ based on $m_\theta(Y_i, Y_i^0, X_i)$ as in Section 6. The results of the simulations are summarized in Table 3 and Table 4.

Similarly to the AR(3) example, both the transition parameters and the slope parameters of $\hat{\theta}^a$ are negatively biased for the three sample sizes under consideration. This is particularly true for the "between" state-dependence parameters $\hat{\gamma_{12}}^a, \hat{\gamma_{21}}^a$ which maintain a small bias even for $N = 8000, 16000$. By comparison, the rescaled GMM estimator $\hat{\theta}^b$ and the iterated

Table 3: Performance of GMM estimators for the bivariate VAR(1): transition parameters

		$\hat{\gamma_{11}}^a$	$\hat{\gamma_{11}}^b$	$\hat{\gamma_{11}}^c$	$\hat{\gamma_{12}}^a$	$\hat{\gamma_{12}}^b$	$\hat{\gamma_{12}}^c$	$\hat{\gamma_{21}}^a$	$\hat{\gamma_{21}}^b$	$\hat{\gamma_{21}}^c$	$\hat{\gamma_{22}}^a$	$\hat{\gamma_{22}}^b$	$\hat{\gamma_{22}}^c$
$N = 2000$													
	Bias	-0.23	0.10	-0.05	-0.21	-0.04	-0.04	-0.20	-0.06	-0.05	-0.24	0.10	-0.05
	MAE	0.27	0.23	0.16	0.29	0.24	0.19	0.27	0.23	0.19	0.27	0.23	0.16
	Iter			5			5			5			5
$N = 8000$													
	Bias	-0.07	0.03	-0.00	-0.08	0.00	-0.00	-0.09	-0.01	-0.01	-0.06	0.03	-0.00
	MAE	0.13	0.11	0.08	0.14	0.12	0.09	0.15	0.12	0.09	0.12	0.11	0.07
	Iter			4			4			4			4
$N = 16000$													
	Bias	-0.04	0.01	-0.00	-0.05	-0.01	-0.00	-0.07	-0.01	-0.00	-0.03	0.01	0.00
	MAE	0.09	0.08	0.05	0.11	0.07	0.06	0.11	0.08	0.06	0.08	0.08	0.06
	Iter			3			3			3			3

NOTES: *Reported results are based on a 1000 replications of the DGP. Bias and MAE stand for median bias and median absolute error respectively. The convergence criterion for the iterated GMM estimator is* $\left\|\hat{\theta}_{s+1} - \hat{\theta}_s\right\| < 10^{-4}$ *and Iter corresponds to the median number of iterations to reach convergence. Bias and MAE for the iterated GMM are reported for replications where convergence is attained which is* $\approx 91\%$ *for* $N = 2000$ *and* $\approx 100\%$ *for* $N = 8000, 16000$.

GMM estimator $\hat{\theta}^c$ demonstrate better accuracy, especially for γ_{12} and γ_{21} which are really the key parameters in our empirical application presented in Section 6. In this specific simulation design, $\hat{\theta}^c$ slightly outperforms $\hat{\theta}^b$ for all $N = 2000, 8000, 16000$ in terms of median bias and median absolute error for the transition parameters. The comparison is somewhat less clear for the slope parameters β_1, β_2. [20]

Surprisingly, when experimenting with a trivariate logit extension, we found that the analog of $\hat{\theta}^b$ performs very poorly for the same simulation design relative to the iterated GMM estimator or even the naive equally-weighted GMM estimator $\hat{\theta}^a$. This is perhaps due to the "large" rescaling factor applied to each valid moment function in that case which pose problems for the optimization of the GMM objective. We have not investigated these peculiarities - which could be design specific - further at this moment but a more thorough analysis of the behavior of GMM in future work would be beneficial. The good performance of $\hat{\theta}^c$ and this shortcoming of $\hat{\theta}^b$ in the trivariate case was one additional motivation for

[20] We also experimented with an iterated GMM estimator based on $\widetilde{m_{\theta}}(Y_i, Y_i^0, X_i)$ and found nearly identical results to $\hat{\theta}^b$.

Table 4: Performance of GMM estimators for the bivariate VAR(1): slope parameters

		$\hat{\beta}_1^{\,a}$	$\hat{\beta}_1^{\,b}$	$\hat{\beta}_1^{\,c}$	$\hat{\beta}_2^{\,a}$	$\hat{\beta}_2^{\,b}$	$\hat{\beta}_2^{\,c}$
$N = 2000$							
	Bias	-0.04	0.01	-0.01	-0.04	0.00	-0.01
	MAE	0.06	0.06	0.06	0.06	0.06	0.05
	Iter			5			5
$N = 8000$							
	Bias	-0.01	-0.00	0.00	-0.01	0.00	0.00
	MAE	0.03	0.03	0.03	0.03	0.03	0.03
	Iter			4			4
$N = 16000$							
	Bias	-0.00	0.00	0.01	-0.00	0.00	0.01
	MAE	0.02	0.02	0.02	0.02	0.02	0.02
	Iter			3			3

NOTES: *Reported results are based on a 1000 replications of the DGP. Bias and MAE stand for median bias and median absolute error respectively. The convergence criterion for the iterated GMM estimator is* $\left\|\hat{\theta}_{s+1} - \hat{\theta}_s\right\| < 10^{-4}$ *and Iter corresponds to the median number of iterations to reach convergence. Bias and MAE for the iterated GMM are reported for replications where convergence is attained which is* $\approx 91\%$ *for* $N = 2000$ *and* $\approx 100\%$ *for* $N = 8000, 16000$.

concentrating on the iterated GMM estimator in our empirical application.

E Proofs of Theorem 1 and Theorem 3

We focus our attention on proving Theorem 3 since proving Theorem 1 would follow nearly identical arguments. At each important step of the proof, we highlight where the arguments for the AR(1) would differ.

Fix a history $y \in \mathcal{Y}^T$ and consider the corresponding basis element $\mathbb{1}\{. = y\}$ of $\mathbb{R}^{\mathcal{Y}^T}$. We have:

$$\mathcal{E}^{(p)}_{y^0,x}\left[\mathbb{1}\{. = y\}\right] = P(Y_i = y | Y_i^0 = y^0, X_i = x, A_i = .)$$

where by definition, for all $a \in \mathbb{R}$,

$$P(Y_i = y|Y_i^0 = y^0, X_i = x, A_i = a) = \frac{N^{y|y^0}(e^a)}{D^{y|y^0}(e^a)}$$
$$N^{y|y^0}(e^a) = \prod_{t=1}^{T} e^{y_t\left(\sum_{r=1}^{p} \gamma_{0r} y_{t-r} + x_t'\beta_0 + a\right)}$$
$$D^{y|y^0}(e^a) = \prod_{t=1}^{T} \left(1 + e^{\sum_{r=1}^{p} \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}\right)$$

Notice that $N^{y|y^0}(e^a)$ and $D^{y|y^0}(e^a)$ are just polynomials of e^a - with dependence on x suppressed for conciseness - and that we always have $\deg\left(N^{y|y^0}(e^a)\right) \leq \deg\left(D^{y|y^0}(e^a)\right)$ with strict inequality unless $y = 1_T$. Moreover, since by assumption for any $t, s \in \{1, \ldots, T-1\}$ and $y, \tilde{y} \in \mathcal{Y}^p$, $\gamma_0' y + x_t'\beta_0 \neq \gamma_0'\tilde{y} + x_s'\beta_0$ if $t \neq s$ or $y \neq \tilde{y}$, $D^{y|y^0}(e^a)$ is a product of distinct irreducible polynomials in e^a. Therefore, by standard results on *partial fraction decompositions*, we know that there exists a unique set of coefficients $(\lambda_0^y, \lambda_1^y, \ldots, \lambda_T^y) \in \mathbb{R}^{T+1}$ independent of the fixed effect such that:

$$P(Y_i = y|Y_i^0 = y^0, X_i = x, A_i = a) = \lambda_0^y + \sum_{t=1}^{T} \lambda_t^y \frac{1}{1 + e^{\sum_{r=1}^{p} \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}}$$
$$= \lambda_0^y + T_0(a) + T_1(a) + T_2(a)$$
$$T_0(a) = \lambda_1^y \frac{1}{1 + e^{\sum_{r=1}^{p} \gamma_{0r} y_{1-r} + x_1'\beta_0 + a}}$$
$$T_1(a) = \sum_{t=2}^{p} \lambda_t^y \frac{1}{1 + e^{\sum_{r=1}^{p} \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}}$$
$$T_3(a) = \sum_{t=p+1}^{T} \lambda_t^y \frac{1}{1 + e^{\sum_{r=1}^{p} \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}}$$

with $\lambda_0^y = 0$ unless $y = 1_T$. This decomposition breaks down the conditional probability $P(Y_i = y|Y_i^0 = y^0, X_i = x, A_i = a)$ into components that depend on the initial condition, namely $T_0(a), T_1(a)$, and components that do not, i.e $T_2(a)$. Notice that $T_1(a)$ would not appear in the AR(1) case. Starting with the first group, we can write:

$$\begin{aligned}
T_0(a) &= \lambda_1^y \pi_0^{0|y^0}(a,x) \\
&= \lambda_1^y \mathbb{1}\{y_0 = 0\}\pi_0^{y_0|y^0}(x,a) + \lambda_1^y \mathbb{1}\{y_0 = 1\}\left(1 - \pi_0^{y_0|y^0}(x,a)\right) \\
&= \lambda_1^y \mathbb{1}\{y_0 = 1\} + \lambda_1^y \mathbb{1}\{y_0 = 0\}\pi_0^{y_0|y^0}(x,a) - \lambda_1^y \mathbb{1}\{y_0 = 1\}\pi_0^{y_0|y^0}(x,a)
\end{aligned}$$

and

$$\begin{aligned}
T_1(a) = &\sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_1^{t-1} \in \mathcal{Y}^{t-1}} \mathbb{1}\{y_{t-1} = \tilde{y}_1, \ldots, y_1 = \tilde{y}_{t-1}\}\pi_{t-1}^{0|\tilde{y}_1^{t-1}, y_0, \ldots, y_{-(p-t)}}(a,x) \\
= &\sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_2^{t-2} \in \mathcal{Y}^{t-2}} \mathbb{1}\{y_{t-1} = 0, y_{t-2} = \tilde{y}_2, \ldots, y_1 = \tilde{y}_{t-1}\}\pi_{t-1}^{0|0, \tilde{y}_2^{t-1}, y_0, \ldots, y_{-(p-t)}}(a,x) \\
+ &\sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_2^{t-2} \in \mathcal{Y}^{t-2}} \mathbb{1}\{y_{t-1} = 1, y_{t-2} = \tilde{y}_2, \ldots, y_1 = \tilde{y}_{t-1}\}\left(1 - \pi_{t-1}^{1|1, \tilde{y}_2^{t-1}, y_0, \ldots, y_{-(p-t)}}(a,x)\right) \\
= &\sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_2^{t-2} \in \mathcal{Y}^{t-2}} \mathbb{1}\{y_{t-1} = 1, y_{t-2} = \tilde{y}_2, \ldots, y_1 = \tilde{y}_{t-1}\} \\
+ &\sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_2^{t-2} \in \mathcal{Y}^{t-2}} \mathbb{1}\{y_{t-1} = 0, y_{t-2} = \tilde{y}_2, \ldots, y_1 = \tilde{y}_{t-1}\}\pi_{t-1}^{0|0, \tilde{y}_2^{t-1}, y_0, \ldots, y_{-(p-t)}}(a,x) \\
- &\sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_2^{t-2} \in \mathcal{Y}^{t-2}} \mathbb{1}\{y_{t-1} = 1, y_{t-2} = \tilde{y}_2, \ldots, y_1 = \tilde{y}_{t-1}\}\pi_{t-1}^{1|1, \tilde{y}_2^{t-1}, y_0, \ldots, y_{-(p-t)}}(a,x)
\end{aligned}$$

Then, for the second group,

$$
\begin{aligned}
T_3(a) &= \sum_{t=p+1}^{T} \lambda_t^{y,y^0} \sum_{\tilde{y}_1^p \in \mathcal{Y}^p} \mathbb{1}\{y_{t-1} = \tilde{y}_1, \ldots, y_{t-p} = \tilde{y}_p\} \pi_{t-1}^{0|\tilde{y}_1^p}(a,x) \\
&= \sum_{t=p+1}^{T} \lambda_t^{y,y^0} \sum_{\tilde{y}_2^p \in \mathcal{Y}^{p-1}} \mathbb{1}\{y_{t-1} = 0, y_{t-2} = y_2, \ldots, y_{t-p} = \tilde{y}_p\} \pi_{t-1}^{0|0,\tilde{y}_2^p}(a,x) \\
&+ \sum_{t=p+1}^{T} \lambda_t^{y,y^0} \sum_{\tilde{y}_2^{p-1} \in \mathcal{Y}^{p-1}} \mathbb{1}\{y_{t-1} = 1, y_{t-2} = y_2, \ldots, y_{t-p} = \tilde{y}_p\} \left(1 - \pi_{t-1}^{1|1,\tilde{y}_2^p}(a,x)\right) \\
&= + \sum_{t=p+1}^{T} \lambda_t^{y,y^0} \sum_{\tilde{y}_2^{p-1} \in \mathcal{Y}^{p-1}} \mathbb{1}\{y_{t-1} = 1, y_{t-2} = y_2, \ldots, y_{t-p} = \tilde{y}_p\} \\
&+ \sum_{t=p+1}^{T} \lambda_t^{y,y^0} \sum_{\tilde{y}_2^p \in \mathcal{Y}^{p-1}} \mathbb{1}\{y_{t-1} = 0, y_{t-2} = y_2, \ldots, y_{t-p} = \tilde{y}_p\} \pi_{t-1}^{0|0,\tilde{y}_2^p}(a,x) \\
&- \sum_{t=p+1}^{T} \lambda_t^{y,y^0} \sum_{\tilde{y}_2^{p-1} \in \mathcal{Y}^{p-1}} \mathbb{1}\{y_{t-1} = 1, y_{t-2} = y_2, \ldots, y_{t-p} = \tilde{y}_p\} \pi_{t-1}^{1|1,\tilde{y}_2^p}(a,x)
\end{aligned}
$$

The unique decompositions for each term make it clear that

$$
\mathcal{F}_{y^0,p,T} = \left\{ 1, \pi_0^{y_0|y^0}(.,x), \left\{ \left(\pi_{t-1}^{y_1|y_1^{t-1},y_0,\ldots,y_{-(p-t)}}(.,x)) \right)_{y_1^{t-1} \in \mathcal{Y}^{t-1}} \right\}_{t=2}^{p}, \left\{ \left(\pi_{t-1}^{y_1|y_1^p}(.,x) \right)_{y_1^p \in \mathcal{Y}^p} \right\}_{t=p+1}^{T} \right\}
$$

forms a basis of $\mathrm{Im}\left(\mathcal{E}_{y^0,x}^{(p)}\right)$ if we can show that the transition probabilities are elements of $\mathrm{Im}\left(\mathcal{E}_{y^0,x}^{(p)}\right)$. We now argue that it is indeed the case:

- First, $\pi_0^{y_0|y^0}(.,x) \in \mathrm{Im}\left(\mathcal{E}_{y^0,x}^{(p)}\right)$ since

$$
\begin{aligned}
\mathbb{E}[(1 - Y_{i1})|Y_i^0 = y^0, X_i = x, A_i = a] &= \frac{1}{1 + e^{\sum_{r=1}^p \gamma_{0r} y_{1-r} + x_1' \beta_0 + a}} = \pi_0^{y_0|y^0}(a,x), \quad \text{if } y_0 = 0 \\
\mathbb{E}[Y_{i1}|Y_i^0 = y^0, X_i = x, A_i = a] &= \frac{e^{\sum_{r=1}^p \gamma_{0r} y_{1-r} + x_1' \beta_0 + a}}{1 + e^{\sum_{r=1}^p \gamma_{0r} y_{1-r} + x_1' \beta_0 + a}} = \pi_0^{y_0|y^0}(a,x), \quad \text{if } y_0 = 1
\end{aligned}
$$

- Second, $\left\{ \left(\pi_{t-1}^{y_1|y_1^p}(.,x) \right)_{y_1^p \in \mathcal{Y}^p} \right\}_{t=p+1}^{T} \in \mathrm{Im}\left(\mathcal{E}_{y^0,x}^{(p)}\right)$ by Theorem 4. For the AR(1) model, one would appeal to Lemma 2.

- Finally, one can easily adapt the proof of Theorem 4 to show that

$\left\{ \left(\pi_{t-1}^{y_1|y_1^{t-1}, y_0, \ldots, y_{-(p-t)}}(., x) \right) \right)_{y_1^{t-1} \in \mathcal{Y}^{t-1}} \right\}_{t=2}^{p} \in \text{Im}\left(\mathcal{E}_{y^0,x}^{(p)} \right)$. First, it follows immediately from Lemma 11 that:

$$\left(\pi_1^{y_1|y_1, y_0, \ldots, y_{-(p-2)}}(., x) \right) \right)_{y_1 \in \mathcal{Y}^{t-1}} \in \text{Im}\left(\mathcal{E}_{y^0,x}^{(p)} \right)$$

Then, by inspecting the induction argument of Theorem 4, it is easily seen that the result that for $T \geq p+1$ and $t \in \{p, \ldots, T-1\}$

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{k+1}}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i) | Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i \right] = \pi_t^{y_1|y_1^{k+1}, Y_{it-(k+1)}, \ldots, Y_{it-(p-1)}}(A_i, X_i)$$

for $k = 0, \ldots, p-2$ can be generalized. It actually holds for $t = k+1$ when $k = 0, \ldots, p-2$, yielding

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^t}(Y_{it+1}, Y_{it}, Y_{i1-p}^{t-1}, X_i) | Y_i^0, X_i, A_i \right] = \pi_t^{y_1|y_1^t, Y_{i0}, \ldots, Y_{it-(p-1)}}(A_i, X_i)$$

This is the desired result. The terms $\left\{ \left(\pi_{t-1}^{y_1|y_1^{t-1}, y_0, \ldots, y_{-(p-t)}}(., x) \right) \right)_{y_1^{t-1} \in \mathcal{Y}^{t-1}} \right\}_{t=2}^{p}$ are not present in the AR(1) case which simplifies the argument.

Thus, we have shown that $\mathcal{F}_{y^0,p,T}$ is a basis of $\text{Im}\left(\mathcal{E}_{y^0,x}^{(p)} \right)$. Next, since $\mathcal{E}_{y^0,x}^{(p)}$ is a linear mapping, we know by the *rank nullity theorem* that:

$$\dim\left(\ker(\mathcal{E}_{y^0,x}^{(p)}) \right) = \dim\left(\mathbb{R}^{\{0,1\}^T} \right) - \text{rank}\left(\mathcal{E}_{y^0,x}^{(p)} \right)$$

Therefore, we have the following implications:

1. If $T \leq p$, $|\mathcal{F}_{y^0,p,T}| = 1 + 1 + \sum_{t=2}^{T} 2^{t-1} = 2 + \sum_{t=1}^{T-1} 2^t = 2 + 2\frac{1-2^{T-1}}{1-2} = 2^T$. Hence, $\text{rank}\left(\mathcal{E}_{y^0,x}^{(p)} \right) = 2^T$ and the rank nullity theorem implies $\dim\left(\ker(\mathcal{E}_{y^0,x}^{(p)}) \right) = 0$

2. If $T = p+1$, $|\mathcal{F}_{y^0,p,T}| = 1 + 1 + \sum_{t=2}^{p} 2^{t-1} + 2^p = 2 \times 2^p = 2^{p+1}$. Then, $\text{rank}\left(\mathcal{E}_{y^0,x}^{(p)} \right) = 2^T$ and the rank nullity theorem implies $\dim\left(\ker(\mathcal{E}_{y^0,x}^{(p)}) \right) = 0$

3. If $T \geq p+2$, $|\mathcal{F}_{y^0,p,T}| = 1 + 1 + \sum_{t=2}^{p} 2^{t-1} + 2^p(T-p) = 2^p + 2^p(T-p) = (T-p+1)2^p$. It follows that $\text{rank}\left(\mathcal{E}_{y^0,x}^{(p)} \right) = (T-p+1)2^p$ and $\dim\left(\ker(\mathcal{E}_{y^0,x}^{(p)}) \right) = 2^T - (T-p+1)2^p$

F Proofs of Propositions 1, 2, 4

Propositions 1, 2 and 4 all follow from the same strategy proof based on the the law of iterated expectations. We focus on Proposition 1 here and leave the other cases to the reader.

Take any t, s verifying $T-1 \geq t > s \geq 1$. For any $k \in \mathcal{Y}$, we have

$$\begin{aligned}
\mathbb{E}\left[\psi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s+1})|Y_{i0}, Y_{i1}^{s-1}, A_i\right] &= \mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}) - \phi_{\theta_0}^{k|k}(Y_{is+1}, Y_{is}, Y_{is-1})|Y_{i0}, Y_{i1}^{s-1}, A_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1})|Y_{i0}, Y_{i1}^{t-1}, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, A_i\right] - \pi^{k|k}(A_i) \\
&= \mathbb{E}\left[\pi^{k|k}(A_i)|Y_{i0}, Y_{i1}^{s-1}, A_i\right] - \pi^{k|k}(A_i) \\
&= \pi^{k|k}(A_i) - \pi^{k|k}(A_i) \\
&= 0
\end{aligned}$$

The second and third equalities follow from the law of iterated expectation and Lemma 1.

G Proofs of Lemma 1 and Lemma 2

Without loss of generality, we will consider the case with covariates. The proposed functional form for the transition function $\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)$ implies that it is null when $Y_{it} \neq 0$. Hence

$$\mathbb{E}\left[\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \frac{1}{1+e^{\gamma_0 Y_{it-1}+X_{it}'\beta_0+A_i}} \times \\
\left(\frac{e^{X_{it+1}'\beta_0+A_i}}{1+e^{X_{it+1}'\beta_0+A_i}}\phi_\theta^{0|0}(1,0,Y_{it-1},X_i) + \frac{1}{1+e^{X_{it+1}'\beta_0+A_i}}\phi_\theta^{0|0}(0,0,Y_{it-1},X_i)\right)$$

Thus, to obtain the transition probability $\pi_t^{0|0}(A_i, X_i) = \frac{1}{1+e^{X_{it+1}'\beta_0+A_i}}$ at $\theta = \theta_0$, we must set:

$$\begin{aligned}
&\phi_\theta^{0|0}(1,0,Y_{it-1},X_i) = e^{\gamma Y_{it-1}+(X_{it}-X_{it+1})'\beta} \\
&\phi_\theta^{0|0}(0,0,Y_{it-1},X_i) = 1 \\
&\phi_\theta^{0|0}(k,1,Y_{it-1},X_i) = 0, \quad \forall k \in \mathcal{Y}
\end{aligned}$$

This can be expressed compactly as: $\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = (1-Y_{it})e^{Y_{it+1}\left(\gamma Y_{it-1}-\Delta X_{it+1}'\beta\right)}$

Likewise, for $\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)$ we have:

$$\mathbb{E}\left[\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \frac{e^{\gamma_0 Y_{it-1}+X_{it}'\beta_0+A_i}}{1+e^{\gamma_0 Y_{it-1}+X_{it}'\beta_0+A_i}} \times$$
$$\left(\frac{e^{\gamma_0+X_{it+1}'\beta_0+A_i}}{1+e^{\gamma_0+X_{it+1}'\beta_0+A_i}}\phi_\theta^{1|1}(1,1,Y_{it-1},X_i) + \frac{1}{1+e^{\gamma_0+X_{it+1}'\beta_0+A_i}}\phi_\theta^{1|1}(0,1,Y_{it-1},X_i)\right)$$

Hence, to get $\pi_t^{1|1}(A_i, X_i) = \frac{e^{\gamma_0+X_{it+1}'\beta_0+A_i}}{1+e^{\gamma_0+X_{it+1}'\beta_0+A_i}}$ at $\theta = \theta_0$, we must set:

$$\begin{aligned}
&\phi_\theta^{1|1}(1,1,Y_{it-1},X_i) = 1 \\
&\phi_\theta^{1|1}(0,1,Y_{it-1},X_i) = e^{\gamma(1-Y_{it-1})+(X_{it+1}-X_{it})'\beta} \\
&\phi_\theta^{1|1}(k,0,Y_{it-1},X_i) = 0, \quad \forall k \in \mathcal{Y}
\end{aligned}$$

This can be written succinctly as: $\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = Y_{it}e^{(1-Y_{it+1})\left(\gamma(1-Y_{it-1})+\beta\Delta X_{it+1}\right)}$

H Proofs of Lemmas 3,10 and Corollaries 3.1, 10.1

The proofs of Lemma 3, Lemma 10, Corollary 3.1, Corollary 10.1 all follow the same logic based on the use of a *partial fraction expansion*. We prove Lemma 3 here and leave the other cases to the reader.

The result hinges on the simple rational fraction identity provided in Lemma 8 that for any three reals v, u, a, we have:

$$\begin{aligned}
\frac{1}{1+e^{v+a}} + (1-e^{u-v})\frac{e^{v+a}}{(1+e^{v+a})(1+e^{u+a})} &= \frac{1}{(1+e^{u+a})} \\
\frac{e^{v+a}}{1+e^{v+a}} + (1-e^{-(u-v)})\frac{e^{u+a}}{(1+e^{v+a})(1+e^{u+a})} &= \frac{e^{u+a}}{(1+e^{u+a})}
\end{aligned}$$

By construction for $T \geq 3$, and t, s such that $T-1 \geq t > s \geq 1$:

$$
\begin{aligned}
&\mathbb{E}\left[\zeta_{\theta_0}^{0|0}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \mathbb{E}\left[(1-Y_{is}) + \omega_{t,s}^{0|0}(\theta_0) Y_{is} \phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{1}{1+e^{\mu_s(\theta_0)+A_i}} + \omega_{t,s}^{0|0}(\theta_0)\mathbb{E}\left[Y_{is}\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{1}{1+e^{\mu_s(\theta_0)+A_i}} + \omega_{t,s}^{0|0}(\theta_0)\mathbb{E}\left[Y_{is}|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right]\frac{1}{1+e^{\kappa_t^{0|0}(\theta_0)+A_i}} \\
&= \frac{1}{1+e^{\mu_s(\theta_0)+A_i}} + (1-e^{\kappa_t^{0|0}(\theta_0)-\mu_s(\theta_0)})\frac{e^{\mu_s(\theta_0)+A_i}}{(1+e^{\mu_s(\theta_0)+A_i})(1+e^{\kappa_t^{0|0}(\theta_0)+A_i})} \\
&= \frac{1}{1+e^{\kappa_t^{0|0}(\theta_0)+A_i}} \\
&= \pi_t^{0|0}(A_i, X_i)
\end{aligned}
$$

The second equality follows from the measureability of the weight $\omega_{t,s}^{0|0}(\theta_0)$ with respect to the conditioning set. The third equality follows from the law of iterated expectations and Lemma 2. The penultimate equality uses the first mathematical identity presented above. Similarly,

$$
\begin{aligned}
&\mathbb{E}\left[\zeta_{\theta_0}^{1|1}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \mathbb{E}\left[Y_{is} + \omega_{t,s}^{1|1}(\theta_0)(1-Y_{is}) \phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{e^{\mu_s(\theta_0)+A_i}}{1+e^{\mu_s(\theta_0)+A_i}} + \omega_{t,s}^{1|1}(\theta_0)\mathbb{E}\left[(1-Y_{is})\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{e^{\mu_s(\theta_0)+A_i}}{1+e^{\mu_s(\theta_0)+A_i}} + \omega_{t,s}^{1|1}(\theta_0)\mathbb{E}\left[(1-Y_{is})|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right]\frac{e^{\kappa_t^{1|1}(\theta_0)+A_i}}{1+e^{\kappa_t^{1|1}(\theta_0)+A_i}} \\
&= \frac{e^{\mu_s(\theta_0)+A_i}}{1+e^{\mu_s(\theta_0)+A_i}} + \left(1-e^{-(\kappa_t^{1|1}(\theta_0)-\mu_s(\theta_0))}\right)\frac{e^{\kappa_t^{1|1}(\theta_0)+A_i}}{(1+e^{\mu_s(\theta_0)+A_i})(1+e^{\kappa_t^{1|1}(\theta_0)+A_i})} \\
&= \frac{e^{\kappa_t^{1|1}(\theta_0)+A_i}}{1+e^{\kappa_t^{1|1}(\theta_0)+A_i}} \\
&= \pi_t^{1|1}(A_i, X_i)
\end{aligned}
$$

The second equality follows from the measurability of the weight $\omega_{t,s}^{0|0}(\theta_0)$ with respect to the conditioning set. The third equality follows from the law of iterated expectations and Lemma 2. The penultimate equality uses the second mathematical identity presented above.

I Proof of Theorem 4

We start by proving the following Lemma

Lemma 11. *In model (5), with $T \geq 2$ and $t \in \{1, \dots, T-1\}$, let*

$$\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) = (1 - Y_{it})e^{Y_{it+1}(\gamma_1 Y_{it-1} - \sum_{l=2}^p \gamma_l \Delta Y_{it+1-l} - \Delta X_{it+1}' \beta)}$$
$$\phi_\theta^{1|1}(Y_{it+1}, Y_t, Y_{it-p}^{t-1}, X_i) = Y_{it} e^{(1-Y_{it+1})\left(\gamma_1(1-Y_{it-1}) + \sum_{l=2}^p \gamma_l \Delta Y_{it+1-l} + \Delta X_{it+1}' \beta\right)}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) | Y_i^0, Y_{i1}^{t-1}, X_i, A_i\right] &= \pi_t^{0|0, Y_{it-1}, \dots, Y_{it-(p-1)}}(A_i, X_i) \\
&= \frac{1}{1 + e^{\sum_{l=2}^p \gamma_{0l} Y_{it+1-l} + X_{it+1}' \beta_0 + A_i}} \\
\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) | Y_i^0, Y_{i1}^{t-1}, X_i, A_i\right] &= \pi_t^{1|1, Y_{it-1}, \dots, Y_{it-(p-1)}}(A_i, X_i) \\
&= \frac{e^{\gamma_{01} + \sum_{l=2}^p \gamma_{0l} Y_{it+1-l} + X_{it+1}' \beta_0 + A_i}}{1 + e^{\gamma_{01} + \sum_{l=2}^p \gamma_{0l} Y_{it+1-l} + X_{it+1}' \beta_0 + A_i}}
\end{aligned}$$

Instead of verifying the result directly from the expression given in the Lemma, it is easier to start from the heuristic idea, emphasized throughout the text, that we look for two functions such that:

$$\begin{aligned}
&\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) = (1 - Y_{it})\phi_\theta^{0|0}(Y_{it+1}, 0, Y_{it-p}^{t-1}, X_i) \\
&\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = Y_{it}\phi_\theta^{1|1}(Y_{it+1}, 1, Y_{it-p}^{t-1}, X_i) \\
&\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) | Y_i^0, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{k|k, Y_{it-1}, \dots, Y_{it-(p-1)}}(A_i, X_i), \quad \forall k \in \mathcal{Y}
\end{aligned}$$

By definition, $\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i)$ is null when $Y_{it} \neq 0$. Hence

$$\mathbb{E}\left[\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) | Y_i^0, Y_{i1}^{t-1}, X, A\right] = \frac{1}{1 + e^{\sum_{l=1}^p \gamma_{0l} Y_{it-l} + X_{it}' \beta_0 + A_i}} \times ($$
$$\frac{e^{\sum_{l=2}^p \gamma_{0l} Y_{it+1-l} + X_{it+1}' \beta_0 + A_i}}{1 + e^{\sum_{l=2}^p \gamma_{0l} Y_{it+1-l} + X_{it+1}' \beta_0 + A_i}} \phi_\theta^{0|0}(1, 0, Y_{it-p}^{t-1}, X_i) + \frac{1}{1 + e^{\gamma_{02} Y_{it-1} + X_{it+1}' \beta_0 + A_i}} \phi_\theta^{0|0}(0, 0, Y_{it-p}^{t-1}, X_i))$$

Thus, to obtain $\pi_t^{0|0,Y_{it-1},\ldots,Y_{it-(p-1)}}(A_i,X_i)=\frac{1}{1+e^{\sum_{l=2}^{p}\gamma_{0l}Y_{it+1-l}+X'_{it+1}\beta_0+A_i}}$ at $\theta=\theta_0$, we must set:

$$\begin{aligned}
&\phi_\theta^{0|0}(1,0,Y_{it-p}^{t-1},X_i)=e^{\gamma_1 Y_{it-1}-\sum_{l=2}^{p}\gamma_l\Delta Y_{it+1-l}-\Delta X'_{it+1}\beta}\\
&\phi_\theta^{0|0}(0,0,Y_{it-p}^{t-1},X_i)=1\\
&\phi_\theta^{0|0}(k,1,Y_{it-p}^{t-1},X_i)=0,\forall k\in\mathcal{Y}
\end{aligned}$$

more compactly this writes,

$$\phi_\theta^{0|0}(Y_{it+1},Y_{it},Y_{it-p}^{t-1},X_i)=(1-Y_{it})e^{Y_{it+1}(\gamma_1 Y_{it-1}-\sum_{l=2}^{p}\gamma_l\Delta Y_{it+1-l}-\Delta X'_{it+1}\beta)}$$

Analogously, $\phi_\theta^{1|1}(Y_{it+1},Y_{it},Y_{it-p}^{t-1},X_i)$ is null when $Y_{it}\neq 1$. Hence

$$\mathbb{E}\left[\phi_\theta^{1|1}(Y_{it+1},Y_{it},Y_{it-p}^{t-1},X_i))|Y_i^0,Y_1^{t-1},X,A\right]=\frac{e^{\sum_{l=1}^{p}\gamma_{0l}Y_{it-l}+X'_{it}\beta_0+A_i}}{1+e^{\sum_{l=1}^{p}\gamma_{0l}Y_{it-l}+X'_{it}\beta_0+A_i}}\times($$
$$\frac{e^{\gamma_{01}+\sum_{l=2}^{p}\gamma_{0l}Y_{it+1-l}+X'_{it+1}\beta_0+A_i}}{1+e^{\gamma_{01}+\sum_{l=2}^{p}\gamma_{0l}Y_{it+1-l}+X'_{it+1}\beta_0+A_i}}\phi_\theta^{1|1}(1,1,Y_{it-p}^{t-1},X_i)+\frac{1}{1+e^{\gamma_{01}+\gamma_{02}Y_{it-1}+X'_{it+1}\beta_0+A_i}}\phi_\theta^{1|1}(0,1,Y_{it-p}^{t-1},X_i))$$

Consequently, to get $\pi_t^{1|1,Y_{it-1},\ldots,Y_{it-(p-1)}}(A_i,X_i)=\frac{e^{\gamma_{01}+\sum_{l=2}^{p}\gamma_{0l}Y_{it+1-l}+X'_{it+1}\beta_0+A_i}}{1+e^{\gamma_{01}+\sum_{l=2}^{p}\gamma_{0l}Y_{it+1-l}+X'_{it+1}\beta_0+A_i}}$ at $\theta=\theta_0$, we must set:

$$\begin{aligned}
&\phi_\theta^{1|1}(1,1,Y_{it-p}^{t-1},X_i)=1\\
&\phi_\theta^{1|1}(0,1,Y_{it-p}^{t-1},X_i)=e^{\gamma_1(1-Y_{it-1})+\sum_{l=2}^{p}\gamma_l\Delta Y_{it+1-l}+\Delta X'_{it+1}\beta}\\
&\phi_\theta^{1|1}(k,0,Y_{it-p}^{t-1},X_i)=0,\forall k\in\mathcal{Y}
\end{aligned}$$

This can be written succinctly as:

$$\phi_\theta^{1|1}(Y_{it+1},Y_t,Y_{it-p}^{t-1},X_i)=Y_{it}e^{(1-Y_{it+1})\left(\gamma_1(1-Y_{it-1})+\sum_{l=2}^{p}\gamma_l\Delta Y_{it+1-l}+\Delta X'_{it+1}\beta\right)}$$

which completes the proof of the Lemma.

Now, for $T\geq p+1$ fix $t\in\{p,\ldots,T-1\}$ and $y=(y_1,\ldots,y_p)=y_1^p\in\{0,1\}^p$. We will prove by finite induction the statement $\mathcal{P}(k)$:

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{k+1}}(Y_{it+1},Y_{it},Y_{it-(p+k)}^{t-1},X_i)|Y_i^0,Y_{i1}^{t-(k+1)},X_i,A_i\right]=\pi_t^{y_1|y_1^{k+1},Y_{it-(k+1)},\ldots,Y_{it-(p-1)}}(A_i,X_i)$$

for $k=0,\ldots,p-2$ for $p\geq 2$.

Base step:

$\mathcal{P}(0)$ is true by Lemma 11 which also deals with the edge case $p = 2$. Thus, let us assume $p \geq 3$ in the remainder of the induction argument.

Induction Step:

Suppose $\mathcal{P}(k-1)$ is true for some $k \in \{1, \dots, p-2\}$, we show that $\mathcal{P}(k)$ is true. Using the law of iterated expectations, the induction hypothesis $\mathcal{P}(k-1)$ and the identities of Lemma 8, we have:

If $y_1 = 0, y_{k+1} = 1$

$$
\begin{aligned}
&\mathbb{E}\left[\phi_{\theta_0}^{0|0,y_2^k,1}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \mathbb{E}\left[(1 - Y_{it-k}) + w_t^{0|0,y_2^k,1}(\theta_0)\phi_{\theta_0}^{0|0,y_2^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)Y_{it-k}|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}} \\
&+ w_t^{0|0,y_2^k,1}(\theta_0)\mathbb{E}\left[\mathbb{E}\left[\phi_{\theta_0}^{0|0,y_2^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-k}, X_i, A_i\right] Y_{it-k}|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}} w_t^{0|0,y_2^k,1}(\theta_0)\mathbb{E}\left[\pi_t^{0|0,y_2^k,Y_{it-k},\dots,Y_{it-(p-1)}}(A_i, X_i)Y_{it-k}|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}} + w_t^{0|0,y_2^k,1}(\theta_0)\mathbb{E}\left[\frac{1}{1 + e^{\sum_{r=2}^{k}\gamma_{0r}y_r + \sum_{r=k+1}^{p}\gamma_{0r}Y_{it-(r-1)} + X_{it+1}'\beta_0 + A_i}} Y_{it-k}|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}} + (1 - e^{(k_t^{0|0,y_2^k,1}(\theta_0) - u_{t-k}(\theta_0))})\frac{1}{1 + e^{k_t^{0|0,y_2^k,1}(\theta_0)+A_i}}\frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} \\
&= \frac{1}{1 + e^{k_t^{0|0,y_2^k,1}(\theta_0)+A_i}} \\
&= \pi_t^{0|0,y_2^k,1,Y_{it-(k+1)},\dots,Y_{it-(p-1)}}(A_i, X_i)
\end{aligned}
$$

If $y_1 = 0, y_{k+1} = 0$

$$
\begin{aligned}
&\mathbb{E}\left[\phi_{\theta_0}^{0|0,y_2^k,0}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \mathbb{E}\left[1 - Y_{it-k} - w_t^{0|0,y_2^k,0}(\theta_0)\left(1 - \phi_{\theta_0}^{0|0,y_2^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)\right)(1 - Y_{it-k})|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= 1 - \frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} \\
&- w_t^{0|0,y_2^k,0}(\theta_0)\times \\
&\mathbb{E}\left[\mathbb{E}\left[\left(1 - \phi_{\theta_0}^{0|0,y_2^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)\right)|Y_i^0, Y_{i1}^{t-k}, X_i, A_i\right](1 - Y_{it-k})|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= 1 - \frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} - w_t^{0|0,y_2^k,0}(\theta_0)\mathbb{E}\left[(1 - \pi_t^{0|0,y_2^k,Y_{it-k},\ldots,Y_{it-(p-1)}}(A_i, X_i))(1 - Y_{it-k})|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= 1 - \frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} \\
&- w_t^{0|0,y_2^k,0}(\theta_0)\mathbb{E}\left[\frac{e^{\sum_{r=2}^k \gamma_{0r} y_r + \sum_{r=k+1}^p \gamma_{0r} Y_{it-(r-1)} + X_{it+1}'\beta_0 + A_i}}{1 + e^{\sum_{r=2}^k \gamma_{0r} y_r + \sum_{r=k+1}^p \gamma_{0r} Y_{it-(r-1)} + X_{it+1}'\beta_0 + A_i}}(1 - Y_{it-k})|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= 1 - \left(\frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} + (1 - e^{-(k_t^{0|0,y_2^k,0}(\theta_0) - u_{t-k}(\theta_0))})\frac{e^{k_t^{0|0,y_2^k,0}(\theta_0)+A_i}}{1 + e^{k_t^{0|0,y_2^k,0}(\theta_0)+A_i}}\frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}}\right) \\
&= 1 - \frac{e^{k_t^{0|0,y_2^k,0}(\theta_0)+A_i}}{1 + e^{k_t^{0|0,y_2^k,0}(\theta_0)+A_i}} \\
&= \frac{1}{1 + e^{k_t^{0|0,y_2^k,0}(\theta_0)+A_i}} \\
&= \pi_t^{0|0,y_2^k,0,Y_{it-(k+1)},\ldots,Y_{it-(p-1)}}(A_i, X_i)
\end{aligned}
$$

If $y_1 = 1, y_{k+1} = 0$

$$
\begin{aligned}
&\mathbb{E}\left[\phi_{\theta_0}^{1|1,y_2^k,0}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \mathbb{E}\left[Y_{it-k} + w_t^{1|1,y_2^k,0}(\theta_0)\phi_{\theta_0}^{1|1,y_2^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)(1 - Y_{it-k})|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} + w_t^{1|1,y_2^k,0}(\theta_0)\times \\
&\mathbb{E}\left[\mathbb{E}\left[\phi_{\theta_0}^{1|1,y_2^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-k}, X_i, A_i\right](1 - Y_{it-k})|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} + w_t^{1|1,y_2^k,0}(\theta_0)\mathbb{E}\left[\pi_t^{1|1,y_2^k,Y_{it-k},\ldots,Y_{it-(p-1)}}(A_i, X_i)(1 - Y_{it-k})|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} \\
&+ w_t^{1|1,y_2^k,0}(\theta_0)\mathbb{E}\left[\frac{e^{\gamma_{01}+\sum_{r=2}^k \gamma_{0r}y_r + \sum_{r=k+1}^p \gamma_{0r}Y_{it-(r-1)} + X_{it+1}'\beta_0 + A_i}}{1 + e^{\gamma_{01}+\sum_{r=2}^k \gamma_{0r}y_r + \sum_{r=k+1}^p \gamma_{0r}Y_{it-(r-1)} + X_{it+1}'\beta_0 + A_i}}(1 - Y_{it-k})|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}} + (1 - e^{-(k_t^{1|1,y_2^k,0}(\theta_0) - u_{t-k}(\theta_0))})\frac{e^{k_t^{1|1,y_2^k,0}(\theta_0)+A_i}}{1 + e^{k_t^{1|1,y_2^k,0}(\theta_0)+A_i}}\frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}} \\
&= \frac{e^{k_t^{1|1,y_2^k,0}(\theta_0)+A_i}}{1 + e^{k_t^{1|1,y_2^k,0}(\theta_0)+A_i}} \\
&= \pi_t^{1|1,y_2^k,0,Y_{it-(k+1)},\ldots,Y_{it-(p-1)}}(A_i, X_i)
\end{aligned}
$$

If $y_1 = 1, y_{k+1} = 1$

$$
\begin{aligned}
&\mathbb{E}\left[\phi_{\theta_0}^{1|1,y_2^k,1}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= \mathbb{E}\left[1 - (1 - Y_{it-k}) - w_t^{1|1,y_2^k,1}(\theta_0)\left(1 - \phi_{\theta_0}^{1|1,y_2^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)\right) Y_{it-k}|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= 1 - \frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}} \\
&- w_t^{1|1,y_2^k,1}(\theta_0)\mathbb{E}\left[\mathbb{E}\left[\left(1 - \pi_t^{1|1,y_2^k,Y_{it-k},\ldots,Y_{it-(p-1)}}(A_i, X_i)\right)|Y_i^0, Y_{i1}^{t-k}, X_i, A_i\right] Y_{it-k}|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= 1 - \frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}} \\
&- w_t^{1|1,y_2^k,1}(\theta_0)\mathbb{E}\left[\frac{1}{1 + e^{\gamma_{01}+\sum_{r=2}^k \gamma_{0r}y_r + \sum_{r=k+1}^p \gamma_{0r}Y_{it-(r-1)} + X_{it+1}'\beta_0 + A_i}} Y_{it-k}|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] \\
&= 1 - \left(\frac{1}{1 + e^{u_{t-k}(\theta_0)+A_i}} + (1 - e^{(k_t^{1|1,y_2^k,1}(\theta_0) - u_{t-k}(\theta_0))})\frac{1}{1 + e^{k_t^{1|1,y_2^k,1}(\theta_0)+A_i}}\frac{e^{u_{t-k}(\theta_0)+A_i}}{1 + e^{u_{t-k}(\theta_0)+A_i}}\right) \\
&= 1 - \frac{1}{1 + e^{k_t^{1|1,y_2^k,1}(\theta_0)+A_i}} \\
&= \frac{e^{k_t^{1|1,y_2^k,1}(\theta_0)+A_i}}{1 + e^{k_t^{1|1,y_2^k,1}(\theta_0)+A_i}} \\
&= \pi_t^{1|1,y_2^k,1,Y_{it-k},\ldots,Y_{it-(p-1)}}(A_i, X_i)
\end{aligned}
$$

Putting these intermediate results together, we have effectively proved that

$$
\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{k+1}}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] = \pi_t^{y_1|y_1^{k+1},Y_{it-(k+1)},\ldots,Y_{it-(p-1)}}(A_i, X_i)
$$

which shows that $\mathcal{P}(k)$ is true and completes the induction argument.

Now, it only remains to show that

$$
\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^p}(Y_{it+1}, Y_{it}, Y_{it-(2p-1)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-p}, X_i, A_i\right] = \pi_t^{y_1|y_1^p}(A_i, X_i)
$$

To this end, it suffices to perform calculations identical to those used in the induction argument but using this time

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{p-1}}(Y_{it+1}, Y_{it}, Y_{it-(2p-2)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-(p-1)}, X_i, A_i\right] = \pi_t^{y_1|y_1^{p-1}, Y_{it-(p-1)}}(A_i, X_i)$$

$$k_t^{y_1|y_1^p}(\theta) = \sum_{r=1}^{p} \gamma_r y_r + X'_{it+1}\beta$$

$$u_{t-(p-1)}(\theta) = \sum_{r=1}^{p} \gamma_r Y_{it-(r+p-1)} + X'_{it-(p-1)}\beta$$

$$w_t^{y_1|y_1^p}(\theta) = \left[1 - e^{(k_t^{y_1|y_1^p}(\theta) - u_{t-(p-1)}(\theta))}\right]^{y_p} \left[1 - e^{-(k_t^{y_1|y_1^p}(\theta) - u_{t-(p-1)}(\theta))}\right]^{1-y_p}$$

This concludes the proof of the theorem.

J Identification of the AR(2) with strictly exogenous regressors

J.1 Identification for $T = 3$ with variability in the initial condition

By Theorem 4, the transition functions associated to: $\pi_2^{0|0,0}(A_i, X_i), \pi_2^{0|0,1}(A_i, X_i), \pi_2^{1|1,0}(A_i, X_i), \pi_2^{1|1,1}(A_i, X_i)$ are given by:

$$\phi_\theta^{0|0,0}(Y_{i3}, Y_{i2}, Y_{i-1}^1, X_i) = e^{\gamma_1 Y_{i0} + \gamma_2 Y_{i-1} - X'_{i31}\beta}(1 - Y_{i1})$$
$$+ \left(1 - e^{\gamma_1 Y_{i0} + \gamma_2 Y_{i-1} - X'_{i31}\beta}\right)(1 - Y_{i1})(1 - Y_{i2})e^{Y_{i3}(\gamma_2 Y_{i0} - X'_{i32}\beta)}$$
$$\phi_\theta^{0|0,1}(Y_{i3}, Y_{i2}, Y_{i-1}^1, X_i) = (1 - Y_{i1}) + \left(1 - e^{-\gamma_1 Y_{i0} + \gamma_2(1-Y_{i-1}) + X'_{i31}\beta}\right) Y_{i1}(1 - Y_{i2})e^{Y_{i3}(\gamma_1 - \gamma_2(1-Y_{i0}) - X'_{i32}\beta)}$$
$$\phi_\theta^{1|1,1}(Y_{i3}, Y_{i2}, Y_{i-1}^1, X_i) = e^{\gamma_1(1-Y_{i0}) + \gamma_2(1-Y_{i-1}) + X'_{i31}\beta}Y_{i1}$$
$$+ \left(1 - e^{\gamma_1(1-Y_{i0}) + \gamma_2(1-Y_{i-1}) + X'_{i31}\beta}\right) Y_{i1}Y_{i2}e^{(1-Y_{i3})(\gamma_2(1-Y_{i0}) + X'_{i32}\beta)}$$
$$\phi_\theta^{1|1,0}(Y_{i3}, Y_{i2}, Y_{i-1}^1, X_i) = Y_{i1} + \left(1 - e^{-\gamma_1(1-Y_{i0}) + \gamma_2 Y_{i-1} - X'_{i31}\beta}\right)(1 - Y_{i1})Y_{i2}e^{(1-Y_{i3})(\gamma_1 - \gamma_2 Y_{i0} + X'_{i32}\beta)}$$

Moreover, an application of Lemma 11 gives

$$\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i-1}^0, X_i) = (1 - Y_{i1})e^{Y_{i2}(\gamma_1 Y_{i0} - \gamma_2(Y_{i0} - Y_{i-1}) - X'_{i21}\beta)}$$
$$\phi_\theta^{1|1}(Y_{i2}, Y_{i1}, Y_{i-1}^0, X_i) = Y_{i1}e^{(1-Y_{i2})\left(\gamma_1(1-Y_{i0}) + \gamma_2(Y_{i0} - Y_{i-1}) + X'_{i21}\beta\right)}$$

such that:

$$\mathbb{E}\left[\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i-1}^0, X_i)|Y_{i-1}, Y_{i0}, A_i\right] = \pi_1^{0|0,Y_{i0}}(A_i, X_i) = \frac{1}{1 + e^{\gamma_2 Y_{i0} + X_{i2}'\beta + A_i}}$$
$$\mathbb{E}\left[\phi_\theta^{1|1}(Y_{i2}, Y_{i1}, Y_{i-1}^0, X_i)|Y_{i-1}, Y_{i0}, A_i\right] = \pi_1^{1|1,Y_{i0}}(A_i, X_i) = \frac{e^{\gamma_1 + \gamma_2 Y_{i0} + X_{i2}'\beta + A_i}}{1 + e^{\gamma_1 + \gamma_2 Y_{i0} + + X_{i2}'\beta + A_i}}$$

For $\pi_2^{0|0,0}(A_i, X_i)$ and $\pi_1^{0|0,Y_{i0}}(A_i, X_i)$ to match, we require both $Y_{i0} = 0$ and $X_{i3} = X_{i2}$ in which case:

$$\begin{aligned}
\phi_\theta^{0|0,0}(Y_{i1}^3, 0, Y_{i-1}, X_i) &= e^{\gamma_2 Y_{i-1} - X_{i31}'\beta}(1 - Y_{i1}) + \left(1 - e^{\gamma_2 Y_{i-1} - X_{i31}'\beta}\right)(1 - Y_{i1})(1 - Y_{i2}) \\
\phi_\theta^{0|0}(Y_{i1}^2, 0, Y_{i-1}, X_i) &= (1 - Y_{i1}) e^{Y_{i2}(\gamma_2 Y_{i-1} - X_{i31}'\beta)} \\
&= (1 - Y_{i1}) Y_{i2} e^{\gamma_2 Y_{i-1} - X_{i31}'\beta} + (1 - Y_{i1})(1 - Y_{i2})
\end{aligned}$$

Therefore,

$$\psi_\theta^{0|0,0}(Y_{i1}^3, 0, Y_{i-1}, X_i) = \phi_\theta^{0|0,0}(Y_{i1}^3, 0, Y_{i-1}, X_i) - \phi_\theta^{0|0}(Y_{i1}^2, 0, Y_{i-1}, X_i) = 0$$

So there is no information about the model parameters in this moment function.

For $\pi_2^{0|0,1}(A_i, X_i)$ and $\pi_1^{0|0,Y_{i0}}(A_i, X_i)$ to match, we require both $Y_{i0} = 1$ and $X_{i3} = X_{i2}$ in which case:

$$\begin{aligned}
\phi_\theta^{0|0,1}(Y_{i1}^3, 1, Y_{i-1}, X_i) &= (1 - Y_{i1}) + \left(1 - e^{-\gamma_1 + \gamma_2(1 - Y_{i-1}) + X_{i31}'\beta}\right) Y_{i1}(1 - Y_{i2}) e^{\gamma_1 Y_{i3}} \\
\phi_\theta^{0|0}(Y_{i1}^2, 1, Y_{i-1}, X_i) &= (1 - Y_{i1}) e^{Y_{i2}(\gamma_1 - \gamma_2(1 - Y_{i-1}) - X_{i31}'\beta)}
\end{aligned}$$

Then, a valid moment condition that depends on all model parameters is:

$$\begin{aligned}
&\psi_\theta^{0|0,1}(Y_{i1}^3, 1, Y_{i-1}, X_i) = \phi_\theta^{0|0,1}(Y_{i1}^3, 1, Y_{i-1}, X_i) - \phi_\theta^{0|0}(Y_{i1}^2, 1, Y_{i-1}, X_i) \\
&= \left(1 - e^{-\gamma_1 + \gamma_2(1 - Y_{i-1}) + X_{i31}'\beta}\right) e^{\gamma_1} Y_{i1}(1 - Y_{i2}) Y_{i3} \\
&+ \left(1 - e^{-\gamma_1 + \gamma_2(1 - Y_{i-1}) + X_{i31}'\beta}\right) Y_{i1}(1 - Y_{i2})(1 - Y_{i3}) \\
&- e^{\gamma_1 - \gamma_2(1 - Y_{i-1}) - X_{i31}'\beta}(1 - e^{-\gamma_1 + \gamma_2(1 - Y_{i-1}) + X_{i31}'\beta})(1 - Y_{i1}) Y_{i2}
\end{aligned}$$

Rescaling this moment function by the factor $\left(e^{\gamma_1 - \gamma_2(1 - Y_{i-1}) - X_{i31}'\beta}(1 - e^{-\gamma_1 + \gamma_2(1 - Y_{i-1}) + X_{i31}'\beta})\right)^{-1}$, one obtains

$$\widetilde{\psi_\theta^{0|0,1}}(Y_{i1}^3, 1, Y_{i-1}, X_i) = e^{\gamma_2(1-Y_{i-1})+X_{i31}'\beta} Y_{i1}(1-Y_{i2})Y_{i3} + e^{-\gamma_1+\gamma_2(1-Y_{i-1})+X_{i31}'\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3})$$
$$- (1-Y_{i1})Y_{i2}$$

Thus, for for the initial condition $Y_{i0} = 1, Y_{i-1} = 1$, we have

$$\widetilde{\psi_\theta^{0|0,1}}(Y_{i1}^3, 1, 1, X_i) = e^{X_{i31}'\beta} Y_{i1}(1-Y_{i2})Y_{i3} + e^{-\gamma_1+X_{i31}'\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3}) - (1-Y_{i1})Y_{i2}$$

which only depends on γ_1 and β. In the notation of Honoré and Weidner (2020), this coincides with their moment function $m_{(1,1)}$. Clearly, it is strictly decreasing in γ_1. Furthermore, this moment function is either increasing or decreasing in β_k depending on the sign of $X_{i3k} - X_{i1k}$. Honoré and Weidner (2020) show that these monotonocity properties can be exploited to uniquely identifies γ_1, β. Instead, for the initial condition $Y_{i0} = 1, Y_{i-1} = 0$, we have

$$\widetilde{\psi_\theta^{0|0,1}}(Y_{i1}^3, 1, 0, X_i) = e^{\gamma_2+X_{i31}'\beta} Y_{i1}(1-Y_{i2})Y_{i3} + e^{-\gamma_1+\gamma_2+X_{i31}'\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3}) - (1-Y_{i1})Y_{i2}$$

which Honoré and Weidner (2020) denote as $m_{(1,0)}$. Provided that γ_1, β are identified, the strict monotonicity of the moment functions in γ_2 ensure that γ_2 is identified.
Analogously, for $\pi_2^{1|1,0}(A_i, X_i)$ and $\pi_1^{0|0,Y_{i0}}(A_i)$ to match, we require both $Y_{i0} = 0$ and $X_{i3} = X_{i2}$ in which case:

$$\begin{aligned}
\phi_\theta^{1|1,0}(Y_{i1}^3, 0, Y_{i-1}, X_i) &= Y_{i1} + \left(1 - e^{-\gamma_1+\gamma_2 Y_{i-1} - X_{i31}'\beta}\right)(1-Y_{i1})Y_{i2} e^{\gamma_1(1-Y_{i3})} \\
\phi_\theta^{1|1}(Y_{i1}^2, 0, Y_{i-1}, X_i) &= Y_{i1} e^{(1-Y_{i2})\left(\gamma_1 - \gamma_2 Y_{i-1} + X_{i31}'\beta\right)}
\end{aligned}$$

Then, a valid moment function that depends on all model parameters is:

$$\begin{aligned}
\psi_\theta^{1|1,0}(Y_{i1}^3, 0, Y_{i-1}, X_i) &= \phi_\theta^{1|1,0}(Y_{i1}^3, 0, Y_{i-1}, X_i) - \phi_\theta^{1|1}(Y_{i1}^2, 0, Y_{i-1}, X_i) \\
&= \left(1 - e^{-\gamma_1+\gamma_2 Y_{i-1} - X_{i31}'\beta}\right) e^{\gamma_1}(1-Y_{i1})Y_{i2}(1-Y_{i3}) \\
&+ \left(1 - e^{-\gamma_1+\gamma_2 Y_{i-1} - X_{i31}'\beta}\right)(1-Y_{i1})Y_{i2}Y_{i3} \\
&- e^{\gamma_1 - \gamma_2 Y_{i-1} + X_{i31}'\beta}\left(1 - e^{-\gamma_1+\gamma_2 Y_{i-1} - X_{i31}'\beta}\right) Y_{i1}(1-Y_{i2})
\end{aligned}$$

Rescaling this moment function by the factor $\left(e^{\gamma_1 - \gamma_2 Y_{i-1} + X_{i31}'\beta}\left(1 - e^{-\gamma_1+\gamma_2 Y_{i-1} - X_{i31}'\beta}\right)\right)^{-1}$, one obtains

$$\widetilde{\psi_\theta^{1|1,0}}(Y_{i1}^3, 0, Y_{i-1}, X_i) = e^{\gamma_2 Y_{i-1} - X_{i31}'\beta}(1-Y_{i1})Y_{i2}(1-Y_{i3}) + e^{-\gamma_1+\gamma_2 Y_{i-1} - X_{i31}'\beta}(1-Y_{i1})Y_{i2}Y_{i3} - Y_{i1}(1-Y_{i2})$$

For the initial condition $Y_{i0} = 0, Y_{i-1} = 0$, we have

$$\widetilde{\psi_\theta^{1|1,0}}(Y_{i1}^3, 0, 0, X_i) = e^{-X_{i31}'\beta}(1-Y_{i1})Y_{i2}(1-Y_{i3}) + e^{-\gamma_1 - X_{i31}'\beta}(1-Y_{i1})Y_{i2}Y_{i3} - Y_{i1}(1-Y_{i2})$$

This moment function also only depends on γ_1, β and coincides with the moment function $m_{(0,0)}$ in Honoré and Weidner (2020). Similarly to $\widetilde{\psi_\theta^{0|0,1}}(Y_{i1}^3, 1, 1, X_i)$, the monotonicity properties of $\widetilde{\psi_\theta^{1|1,0}}(Y_{i1}^3, 0, 0, X_i)$ can be exploited to uniquely identifies γ_1, β (see Honoré and Weidner (2020)). Instead, for the initial condition $Y_{i0} = 0, Y_{i-1} = 1$, we obtain

$$\widetilde{\psi_\theta^{1|1,0}}(Y_{i1}^3, 0, 1, X_i) = e^{\gamma_2 - X_{i31}'\beta}(1-Y_{i1})Y_{i2}(1-Y_{i3}) + e^{-\gamma_1+\gamma_2 - X_{i31}'\beta}(1-Y_{i1})Y_{i2}Y_{i3} - Y_{i1}(1-Y_{i2})$$

Provided that γ_1, β is identified, the strict monotonicity of this moment function in γ_2 implies that it identifies γ_2 uniquely. This is $m_{(0,1)}$ in Honoré and Weidner (2020).

Lastly, for $\pi_2^{1|1,1}(A_i)$ and $\pi_1^{1|1,Y_{i0}}(A_i)$ to match, we require both $Y_{i0} = 1$ and $X_{i3} = X_{i2}$ in which case:

$$\begin{aligned}
\phi_\theta^{1|1,1}(Y_{i1}^3, 1, Y_{i-1}, X_i) &= e^{\gamma_2(1-Y_{i-1}) + X_{i31}'\beta}Y_{i1} + \left(1 - e^{\gamma_2(1-Y_{i-1}) + X_{i31}'\beta}\right)Y_{i1}Y_{i2} \\
\phi_\theta^{1|1}(Y_{i1}^2, 1, Y_{i-1}, X_i) &= Y_{i1}e^{(1-Y_{i2})\left(\gamma_2(1-Y_{i-1}) + X_{i21}'\beta\right)} \\
&= Y_{i1}(1-Y_{i2})e^{\gamma_2(1-Y_{i-1}) + X_{i21}'\beta} + Y_{i1}Y_{i2}
\end{aligned}$$

Then, a valid moment function

$$\begin{aligned}
\psi_\theta^{1|1,1}(Y_{i1}^3, 1, Y_{i-1}, X_i) &= \phi_\theta^{1|1,1}(Y_{i1}^3, 1, Y_{i-1}, X_i) - \phi_\theta^{1|1}(Y_{i1}^2, 1, Y_{i-1}, X_i) \\
&= 0
\end{aligned}$$

is identically zero and hence contains no information about the model parameters.

J.2 Proof of Theorem 5

We recall from the discussion of Section 4.5 that $T = 4$ and $K_x \geq 2$ so that there are at least 2 exogenous explanatory variables. We have $X_{it} = (W_{it}, R_{it}')' \in \mathbb{R}^{K_x}$, $\beta = (\beta_W, \beta_R')' \in \mathbb{R}^{K_x}$

and $Z_i = (R_i', W_{i1}, W_{i3}, W_{i4})' \in \mathbb{R}^{4K_x - 1}$. Our goal is to prove Theorem 5 under Assumptions 2 and 3.

Specializing Proposition 4 to the AR(2) with $T = 4$ yields the valid moment function:

$$
\begin{aligned}
&\psi_\theta^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, X_i) = \left(e^{\gamma_2 Y_{i0} - X_{i42}'\beta} - 1\right)(1 - Y_{i1})(1 - Y_{i2})Y_{i3} \\
&+ \left[e^{\gamma_2 Y_{i0} - X_{i42}'\beta} + \left(1 - e^{\gamma_2 Y_{i0} - X_{i42}'\beta}\right) e^{-X_{i43}'\beta} - 1\right](1 - Y_{i1})(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\
&+ e^{\gamma_1(1 - Y_{i0}) + \gamma_2(Y_{i0} - Y_{i-1}) + X_{i21}'\beta} Y_{i1}(1 - Y_{i2})Y_{i3} \\
&+ e^{-\gamma_1 Y_{i0} - \gamma_2 Y_{i-1} + X_{i41}'\beta}\left[e^{\gamma_1 + \gamma_2 Y_{i0} - X_{i42}'\beta} + \left(1 - e^{\gamma_1 + \gamma_2 Y_{i0} - X_{i42}'\beta}\right) e^{\gamma_2 - X_{i43}'\beta}\right] Y_{i1}(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\
&+ e^{-\gamma_1 Y_{i0} - \gamma_2 Y_{i-1} + X_{i41}'\beta} Y_{i1}(1 - Y_{i2})(1 - Y_{i3})(1 - Y_{i4}) \\
&- (1 - Y_{i1})Y_{i2}
\end{aligned}
$$

Define, the "limiting" moment function, where we have taken W_{i2} to $+\infty$

$$
\begin{aligned}
\psi_{\theta,\infty}^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, Z_i) = &-(1 - Y_{i1})(1 - Y_{i2})Y_{i3} \\
&+ \left[e^{X_{i34}'\beta} - 1\right](1 - Y_{i1})(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\
&+ e^{-\gamma_1 Y_{i0} + \gamma_2(1 - Y_{i-1}) + X_{i31}'\beta} Y_{i1}(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\
&+ e^{-\gamma_1 Y_{i0} - \gamma_2 Y_{i-1} + X_{i41}'\beta} Y_{i1}(1 - Y_{i2})(1 - Y_{i3})(1 - Y_{i4})
\end{aligned} \tag{11}
$$

For $s \in \{-, +\}^{K_x}$, consider the moment objective

$$
\Psi_{s,y^0}^{0|0,0}(\theta) = \lim_{w_2 \to \infty} \mathbb{E}\left[\psi_\theta^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, X_i) | Y_i^0 = y^0, X_i \in \mathcal{X}_s, W_{i2} = w_2\right]
$$

We will show in two successive steps (a) and (b) that

$$
\begin{aligned}
\Psi_{s,y^0}^{0|0,0}(\theta) &= \lim_{w_2 \to \infty} \mathbb{E}\left[\psi_{\theta,\infty}^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, Z_i) | Y_i^0 = y^0, X_i \in \mathcal{X}_s, W_{i2} = w_2\right] \quad \text{(a)} \\
&= \mathbb{E}\left[\psi_{\theta,\infty}^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, Z_i) | Y_i^0 = y^0, X_i \in \mathcal{X}_s, W_{i2} = \infty\right] \quad \text{(b)}
\end{aligned}
$$

To establish (a), we start by observing that the history sequence $(1-Y_{i1})Y_{i2}$ featuring in $\psi_\theta^{0|0,0}$ has expectation zero. To see this, note that by iterated expectations

$$
\begin{aligned}
&\lim_{w_2\to\infty}\mathbb{E}\left[(1-Y_{i1})Y_{i2}|Y_i^0=y^0,X_i\in\mathcal{X}_s,W_{i2}=w_2\right]\\
&=\lim_{w_2\to\infty}\int\frac{e^{\gamma_{02}y_0+x_2'\beta_0+a}}{1+e^{\gamma_{02}y_0+x_2'\beta_0+a}}\frac{1}{1+e^{\gamma_{01}y_0+\gamma_{02}y_{i-1}+x_1'\beta_0+a}}p(a,z|y_0,\mathcal{X}_s,w_2)dadz
\end{aligned}
$$

Now, $p(a,z|y_0,\mathcal{X}_s,w_2)=p(a|y_0,z,w_2)p(z|y_0,\mathcal{X}_s,w_2)=p(a|y_0,z,w_2)\frac{p(z|y_0,w_2)\mathbb{1}\{X_i\in\mathcal{X}_s\}}{\int_{\mathcal{X}_s}p(z|y_0,w_2)dz}$. Hence, by part (iii) of Assumption 3, an integrable dominating function of the integrand is

$$
\frac{e^{\gamma_{02}y_0+x_2'\beta_0+a}}{1+e^{\gamma_{02}y_0+x_2'\beta_0+A_i}}\frac{1}{1+e^{\gamma_{01}y_0+\gamma_{02}y_{i-1}+x_1'\beta_0+a}}p(a,z|y_0,\mathcal{X}_s,w_2)\leq d_0(a)\frac{d_2(z)}{\int_{\mathcal{X}_s}d_1(z)dz}
$$

Moreover, by parts (ii)-(iii) of Assumption 3 and the Dominated Convergence Theorem,

$$
\lim_{w_2\to\infty}p(a,z|y_0,\mathcal{X}_s,w_2)=q(a|y_0,z)\frac{q(z|y_0)\mathbb{1}\{X_i\in\mathcal{X}_s\}}{\int_{\mathcal{X}_s}q(z|y_0)dz}\equiv q(a,z|y_0,\mathcal{X}_s)
$$

Hence another application of the Dominated Convergence Theorem gives

$$
\begin{aligned}
&\lim_{w_2\to\infty}\mathbb{E}\left[(1-Y_{i1})Y_{i2}|Y_i^0=y^0,X_i\in\mathcal{X}_s,W_{i2}=w_2\right]\\
&=\int\lim_{w_2\to\infty}\frac{e^{\gamma_{02}y_0+x_2'\beta_0+a}}{1+e^{\gamma_{02}y_0+x_2'\beta_0+a}}\frac{1}{1+e^{\gamma_{01}y_0+\gamma_{02}y_{i-1}+x_1'\beta_0+a}}p(a,z|y_0,\mathcal{X}_s,w_2)dadz\\
&=\int 0\times q(a,z|y_0,\mathcal{X}_s)dadz\\
&=0
\end{aligned}
$$

where the third line follows from the fact that $\lim_{w_2\to\infty}e^{w_2\beta_W}=0$ by Assumption 2. Applying the same arguments to each remaining summand of $\psi_\theta^{0|0,0}$ and collecting terms delivers (a). To obtain (b), we note that by part (iv) of Assumption 2, $w_2\mapsto\mathbb{E}\left[\psi_{\theta,\infty}^{0|0,0}(Y_{i4},Y_{i3},Y_{i-1}^2,Z_i)|Y_i^0=y^0,X_i\in\mathcal{X}_s,W_{i2}=w_2\right]$ is continuous with a well defined limit at infinity in light of (a). As a result, we can work directly with its continuous extension at infinity.

Let us focus on the initial condition $y_0=y_{-1}=0$. It is clear from Equation (6) that $\Psi_{s,0,0}^{0|0,0}(\theta)$ does not depend on γ_1. Furthermore, by parts (i) of Assumption 3 we note that we have the following integrable dominating functions for the derivative:

$$\left|\frac{\partial \psi^{0|0,0}_{\theta,\infty}(Y_{i4}, Y_{i3}, Y^2_{i-1}, Z_i)}{\partial \gamma_2}\right| = e^{\gamma_2 + X'_{i31}\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3})Y_{i4} \leq \sup_{g_2 \in \mathbb{G}_2, b \in \mathbb{B}} e^{g_2 + 2\max(|\bar{x}|, |\underline{x}|)\|b\|_1}$$

$$\begin{aligned}
\left|\frac{\partial \psi^{0|0,0}_{\theta,\infty}(Y_{i4}, Y_{i3}, Y^2_{i-1}, Z_i)}{\partial \beta_k}\right| &= \Big| X_{ik,34} e^{X'_{i34}\beta}(1-Y_{i1})(1-Y_{i2})(1-Y_{i3})Y_{i4} \\
&\quad + X_{ik,31} e^{\gamma_2 + X'_{i31}\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3})Y_{i4} \\
&\quad + X_{ik,41} e^{\gamma_2 + X'_{i31}\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3})(1-Y_{i4}) \Big| \\
&\leq \big|X_{ik,34}\big| e^{X'_{i34}\beta} + \big|X_{ik,31}\big| e^{\gamma_2 + X'_{i31}\beta} + \big|X_{ik,41}\big| e^{\gamma_2 + X'_{i31}\beta} \\
&\leq 2\max(|\bar{x}|, |\underline{x}|) \sup_{b \in \mathbb{B}} e^{2\max(|\bar{x}|, |\underline{x}|)\|b\|_1}(1 + 2 \sup_{g_2 \in \mathbb{G}_2} e^{g_2})
\end{aligned}$$

Hence, by Leibniz integral rule, we get

$$\begin{aligned}
&\frac{\partial \Psi^{0|0,0}_{s,0,0}(\theta)}{\partial \gamma_2} \\
&= \mathbb{E}\left[\frac{\partial \psi^{0|0,0}_{\theta,\infty}(Y_{i4}, Y_{i3}, Y^2_{i-1}, Z_i)}{\partial \gamma_2} | Y^0_i = (0,0), X_i \in \mathcal{X}_s, W_{i2} = \infty\right] \\
&= \mathbb{E}\left[e^{\gamma_2 + X'_{i31}\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3})Y_{i4} | Y^0_i = (0,0), X_i \in \mathcal{X}_s, W_{i2} = \infty\right] \\
&= \mathbb{E}\left[e^{\gamma_2 + X'_{i31}\beta} \underbrace{\mathbb{E}\left[Y_{i1}(1-Y_{i2})(1-Y_{i3})Y_{i4} | Y^0_i = (0,0), Z_i, W_{i2} = \infty, A_i\right]}_{>0} | Y^0_i = (0,0), X_i \in \mathcal{X}_s, W_{i2} = \infty\right] \\
&> 0
\end{aligned}$$

Similarly,

$$
\begin{aligned}
&\frac{\partial \Psi^{0|0,0}_{s,0,0}(\theta)}{\partial \beta_k} \\
&= \mathbb{E}\left[\frac{\partial \psi^{0|0,0}_{\theta,-\infty}(Y_{i4}, Y_{i3}, Y^2_{i-1}, Z_i)}{\partial \beta_k} | Y^0_i = (0,0), X_i \in \mathcal{X}_s, W_{i2} = \infty\right] \\
&= \mathbb{E}\Big[X_{ik,34} e^{X'_{i34}\beta} \times \\
&\underbrace{\mathbb{E}\left[(1-Y_{i1})(1-Y_{i2})(1-Y_{i3})Y_{i4} | Y^0_i = (0,0), Z_i, W_{i2} = \infty, A_i\right]}_{>0} | Y^0_i = (0,0), X_i \in \mathcal{X}_s, W_{i2} = \infty\Big] \\
&+ \mathbb{E}\Big[X_{ik,31} e^{\gamma_2 + X'_{i31}\beta} \times \\
&\underbrace{\mathbb{E}\left[Y_{i1}(1-Y_{i2})(1-Y_{i3})Y_{i4} | Y^0_i = (0,0), Z_i, W_{i2} = \infty, A_i\right]}_{>0} | Y^0_i = (0,0), X_i \in \mathcal{X}_s, W_{i2} = \infty\Big] \\
&+ \mathbb{E}\Big[X_{ik,41} e^{\gamma_2 + X'_{i31}\beta} \times \\
&\underbrace{\mathbb{E}\left[Y_{i1}(1-Y_{i2})(1-Y_{i3})(1-Y_{i4}) | Y^0_i = (0,0), Z_i, W_{i2} = \infty, A_i\right]}_{>0} | Y^0_i = (0,0), X_i \in \mathcal{X}_s, W_{i2} = \infty\Big]
\end{aligned}
$$

The last display shows that $\frac{\partial \Psi^{0|0,0}_{s,0,0}(\theta)}{\partial \beta_k} > 0$ if $s_k = +$ and $\frac{\partial \Psi^{0|0,0}_{s,0,0}(\theta)}{\partial \beta_k} < 0$ if $s_k = -$. Therefore, appealing to Lemma 2 in Honoré and Weidner (2020), we conclude that the 2^{K_x} system of equations in $K_x + 1$ unkowns given by:

$$
\Psi^{0|0,0}_{s,0,0}(\theta) = 0, \quad \forall s \in \{-,+\}^{K_x}
$$

has at most one solution. It is precisely (γ_{02}, β_0), since the validity of $\psi^{0|0,0}_{\theta}(Y_{i4}, Y_{i3}, Y^2_{i-1}, X_i)$ for arbitrary X_i directly implies the validity of the limiting moment $\psi^{0|0,0}_{\theta,\infty}(Y_{i4}, Y_{i3}, Y^2_{i-1}, Z_i)$ at "$W_{i2} = \infty$". Then, notice that for any other initial condition $y^0 \in \{(0,1),(1,0),(1,1)\}$, the objective $\Psi^{0|0,0}_{s,y^0}(\theta)$ is strictly monotonic in γ_1. Hence, given (γ_{02}, β_0), it point identifies γ_{01}. This concludes the proof of Theorem 5.

K Proof of Proposition 3

We recall that by definition,

$\Pi_t^{k_1^s|l_1^p}(y^0, x_1^{t+s}) =$

$\mathbb{E}\left[P(Y_{it+s} = k_s, \dots, Y_{it+1} = k_1 \,|\, Y_{it} = l_1, \dots, Y_{it-(p-1)} = l_p, X_{i1}^{t+s} = x_1^{t+s}, A_i) \,|\, Y_i^0 = y^0, X_{i1}^{t+s} = x_1^{t+s}\right]$

We have

$$P(Y_{it+s} = k_s, \dots, Y_{it+1} = k_1 \,|\, Y_{it} = l_1, \dots, Y_{it-(p-1)} = l_p, X_{i1}^{t+s} = x_1^{t+s}, A_i) = \frac{N^{k_1^s|l_1^p}(e^a)}{D^{k_1^s|l_1^p}(e^a)}$$

where $N^{k_1^s|l_1^p}(e^a), D^{k_1^s|l_1^p}(e^a)$ are polynomials in e^a. There are two cases to consider.

<u>Case 1:</u> $s < p$

Then,

$$N^{k_1^s|l_1^p}(e^a) = e^{k_1\left(\sum_{r=1}^p \gamma_{0r} l_r + x_{t+1}'\beta_0 + a\right)} \prod_{j=1}^{s-1} e^{k_{j+1}\left(\sum_{r=1}^j \gamma_{0r} k_{j+1-r} + \sum_{r=j+1}^p \gamma_{0r} l_{r-j} + x_{t+1+j}'\beta_0 + a\right)}$$

$$D^{k_1^s|l_1^p}(e^a) = \left(1 + e^{\sum_{r=1}^p \gamma_{0r} l_r + x_{t+1}'\beta_0 + a}\right) \prod_{j=1}^{s-1} \left(1 + e^{\sum_{r=1}^j \gamma_{0r} k_{j+1-r} + \sum_{r=j+1}^p \gamma_{0r} l_{r-j} + x_{t+1+j}'\beta_0 + a}\right)$$

We note that $\deg(N^{k_1^s|l_1^p}(e^a)) \leq \deg(D^{k_1^s|l_1^p}(e^a))$ with strict inequality unless $k_1^s = 1_s$. Furthermore, since by assumption for any $t \in \{p, \dots, T-2\}$, $s \in \{1, \dots, T-1-t\}$ and $y, \tilde{y} \in \mathcal{Y}^p$, $\gamma_0' y + x_t'\beta_0 \neq \gamma_0' \tilde{y} + x_{t+s}'\beta_0$, $D^{k_1^s|l_1^p}(e^a)$ is a product of distinct irreducible polynomials in e^a. Consequently, standard results on *partial fraction decompositions* entail that there exists a unique set of known coefficients $(\mu, \lambda_0, \lambda_1, \dots, \lambda_{s-1}) \in \mathbb{R}^{s+1}$ such that:

$$\frac{N^{k_1^s|l_1^p}(e^a)}{D^{k_1^s|l_1^p}(e^a)} = \mu + \lambda_0 \frac{1}{\left(1 + e^{\sum_{r=1}^p \gamma_{0r} l_r + x_{t+1}'\beta_0 + a}\right)} + \sum_{j=1}^{s-1} \lambda_j \frac{1}{1 + e^{\sum_{r=1}^j \gamma_{0r} k_{j+1-r} + \sum_{r=j+1}^p \gamma_{0r} l_{r-j} + x_{t+1+j}'\beta_0 + a}}$$

with $\mu = 0$ unless $k_1^s = 1_s$. We can rewrite this in terms of transition probabilities as:

$$\begin{aligned}\frac{N^{k_1^s|l_1^p}(e^a)}{D^{k_1^s|l_1^p}(e^a)} &= \mu + \lambda_0 \pi_t^{0|l_1^p}(a, x_{t+1}) + \sum_{j=1}^{s-1} \lambda_j \pi_{t+j}^{0|k_j,\dots,k_1,l_1^{p-j}}(a, x_{t+1+j}) \\ &= \mu + \lambda_0(1-l_1)\pi_t^{l_1|l_1^p}(a, x_{t+1}) + \lambda_0 l_1(1 - \pi_t^{l_1|l_1^p}(a, x_{t+1})) + \\ &\sum_{j=1}^{s-1} \lambda_j (1-k_j)\pi_{t+j}^{k_j|k_j,\dots,k_1,l_1^{p-j}}(a, x_{t+1+j}) + \sum_{j=1}^{s-1} \lambda_j k_j (1 - \pi_{t+j}^{k_j|k_j,\dots,k_1,l_1^{p-j}}(a, x_{t+1+j}))\end{aligned}$$

This last result in conjunction with Theorem 4, implies that:

$$\begin{aligned}\Pi_t^{k_1^s|l_1^p}(y^0, x_1^{t+s}) = \mu \\ &+ \mathbb{E}\Bigg[\lambda_0(1-l_1)\phi_{\theta_0}^{l_1|l_1^p}(Y_{it-(2p-1)}^{t+1}, x_1^{t+s}) + \lambda_0 l_1 \left(1 - \phi_{\theta_0}^{l_1|l_1^p}(Y_{it-(2p-1)}^{t+1}, x_1^{t+s})\right) \\ &+ \sum_{j=1}^{s-1} \lambda_j(1-k_j)\phi_{\theta_0}^{k_j|k_j,\dots,k_1,l_1^{p-j}}(Y_{it+j-(2p-1)}^{t+j+1}, x_1^{t+s}) \\ &+ \sum_{j=1}^{s-1} \lambda_j k_j \left(1 - \phi_{\theta_0}^{k_j|k_j,\dots,k_1,l_1^{p-j}}(Y_{it+j-(2p-1)}^{t+j+1}, x_1^{t+s})\right) \mid Y_i^0 = y^0, X_{i1}^{t+s} = x_1^{t+s}\Bigg]\end{aligned}$$

which shows that $\Pi_t^{k_1^s|l_1^p}(y^0, x_1^{t+s})$ is identified given that θ_0 is identified by assumption.

<u>Case 2:</u> $s \geq p$

Then,

$$\begin{aligned}D^{k_1^s|l_1^p}(e^a) &= \left(1 + e^{\sum_{r=1}^p \gamma_{0r} l_r + x_{t+1}'\beta_0 + a}\right)\prod_{j=1}^{p-1}\left(1 + e^{\sum_{r=1}^j \gamma_{0r}k_{j+1-r} + \sum_{r=j+1}^p \gamma_{0r} l_{r-j} + x_{t+1+j}'\beta_0 + a}\right) \\ &\times \prod_{j=p}^{s-1}\left(1 + e^{\sum_{r=1}^p \gamma_{0r}k_{j+1-r} + x_{t+1+j}'\beta_0 + a}\right)\end{aligned}$$

$$\begin{aligned}N^{k_1^s|l_1^p}(e^a) &= e^{k_1\left(\sum_{r=1}^p \gamma_{0r} l_r + x_{t+1}'\beta_0 + a\right)}\prod_{j=1}^{p-1} e^{k_{j+1}\left(\sum_{r=1}^j \gamma_{0r}k_{j+1-r} + \sum_{r=j+1}^p \gamma_{0r} l_{r-j} + x_{t+1+j}'\beta_0 + a\right)} \\ &\times \prod_{j=p}^{s-1} e^{k_{j+1}\left(\sum_{r=1}^p \gamma_{0r}k_{j+1-r} + x_{t+1+j}'\beta_0 + a\right)}\end{aligned}$$

Invoking identical arguments as in the case $s < p$, there exists a unique set of known coefficients $(\mu, \lambda_0, \lambda_1, \ldots, \lambda_{s-1}) \in \mathbb{R}^{s+1}$ such that:

$$
\begin{aligned}
\Pi_t^{k_1^s|l_1^p}(y^0, x_1^{t+s}) = \mu & \\
& + \mathbb{E}\Bigg[\lambda_0(1-l_1)\phi_{\theta_0}^{l_1|l_1^p}(Y_{it-(2p-1)}^{t+1}, x_1^{t+s}) + \lambda_0 l_1\left(1-\phi_{\theta_0}^{l_1|l_1^p}(Y_{it-(2p-1)}^{t+1}, x_1^{t+s})\right) \\
& + \sum_{j=1}^{p-1}\lambda_j(1-k_j)\phi_{\theta_0}^{k_j|k_j,\ldots,k_1,l_1^{p-j}}(Y_{it+j-(2p-1)}^{t+j+1}, x_1^{t+s}) \\
& + \sum_{j=1}^{p-1}\lambda_j k_j\left(1-\phi_{\theta_0}^{k_j|k_j,\ldots,k_1,l_1^{p-j}}(Y_{it+j-(2p-1)}^{t+j+1}, x_1^{t+s})\right) \\
& + \sum_{j=p}^{s-1}\lambda_j(1-k_j)\phi_{\theta_0}^{k_j|k_j,\ldots,k_{j+1-p}}(Y_{it+j-(2p-1)}^{t+j+1}, x_1^{t+s}) \\
& + \sum_{j=p}^{s-1}\lambda_j k_j\left(1-\phi_{\theta_0}^{k_j|k_j,\ldots,k_{j+1-p}}(Y_{it+j-(2p-1)}^{t+j+1}, x_1^{t+s})\right) \mid Y_i^0 = y^0, X_{i1}^{t+s} = x_1^{t+s}\Bigg]
\end{aligned}
$$

which again shows that $\Pi_t^{k_1^s|l_1^p}(y^0, x_1^{t+s})$ is identified given that θ_0 is identified by assumption. This concludes the proof.

L Proof of Lemma 4

Let

$$
\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = \mathbb{1}\{Y_{it} = k\}e^{\sum_{m=1}^M (Y_{m,it+1}-k_m)\left(\sum_{j=1}^M \gamma_{mj}(Y_{j,it-1}-k_j) - \Delta X'_{m,it+1}\beta_m\right)}
$$

We verify the claim by direct calculation.

$$
\begin{aligned}
&\mathbb{E}\left[\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = P(Y_{it} = k|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i) \\
&\times \sum_{l\in\mathcal{Y}} P(Y_{it+1} = l|Y_{i0}, Y_{i1}^{t-1}, Y_{it} = k, X_i, A_i)\phi_\theta^{k|k}(l, k, Y_{it-1}, X_i) \\
&= \prod_{m=1}^M \frac{e^{k_m(\sum_{j=1}^M \gamma_{mj}Y_{j,it-1} + X'_{m,it}\beta_m + A_{m,i})}}{1 + e^{\sum_{j=1}^M \gamma_{mj}Y_{j,it-1} + X'_{m,it}\beta_m + A_{m,i}}} \\
&\times \sum_{l\in\mathcal{Y}}\prod_{m=1}^M \frac{e^{l_m(\sum_{j=1}^M \gamma_{mj}k_j + X'_{m,it+1}\beta_m + A_{m,i})}}{1 + e^{\sum_{j=1}^M \gamma_{mj}k_j + X'_{m,it+1}\beta_m + A_{m,i}}} e^{\sum_{m=1}^M (l_m-k_m)\left(\sum_{j=1}^M \gamma_{mj}(Y_{j,it-1}-k_j) - \Delta X'_{m,it+1}\beta_m\right)}
\end{aligned}
$$

$$= \sum_{l\in\mathcal{Y}} \prod_{m=1}^{M} \frac{e^{l_m(\sum_{j=1}^{M}\gamma_{mj}Y_{j,it-1}+X'_{m,it}\beta_m+A_{m,i})}}{1+e^{\sum_{j=1}^{M}\gamma_{mj}k_j+X'_{m,it+1}\beta_m+A_{m,i}}} \frac{e^{k_m(\sum_{j=1}^{M}\gamma_{mj}k_j+X'_{m,it+1}\beta_m+A_{m,i})}}{1+e^{\sum_{j=1}^{M}\gamma_{mj}Y_{j,it-1}+X'_{m,it}\beta_m+A_{m,i}}}$$

$$= \prod_{m=1}^{M} \frac{e^{k_m(\sum_{j=1}^{M}\gamma_{mj}k_j+X'_{m,it+1}\beta_m+A_{m,i})}}{1+e^{\sum_{j=1}^{M}\gamma_{mj}Y_{j,it-1}+X'_{m,it}\beta_m+A_{m,i}}} \frac{1}{1+e^{\sum_{j=1}^{M}\gamma_{mj}k_j+X'_{m,it+1}\beta_m+A_{m,i}}} \sum_{l\in\mathcal{Y}} \prod_{m=1}^{M} e^{l_m(\sum_{j=1}^{M}\gamma_{mj}Y_{j,it-1}+X'_{m,it}\beta_m+A_{m,i})}$$

Now, noting that

$$\sum_{l\in\mathcal{Y}} \prod_{m=1}^{M} e^{l_m(\sum_{j=1}^{M}\gamma_{mj}Y_{j,it-1}+X'_{m,it}\beta_m+A_{m,i})} = \prod_{m=1}^{M}(1+e^{\sum_{j=1}^{M}\gamma_{mj}Y_{j,it-1}+X'_{m,it}\beta_m+A_{m,i}})$$

we finally get

$$\mathbb{E}\left[\phi_\theta^{k|k}(Y_{it+1},Y_{it},Y_{it-1},X_i)|Y_{i0},Y_{i1}^{t-1},X_i,A_i\right]$$

$$= \prod_{m=1}^{M} \frac{e^{k_m(\sum_{j=1}^{M}\gamma_{mj}k_j+X'_{m,it+1}\beta_m+A_{m,i})}}{1+e^{\sum_{j=1}^{M}\gamma_{mj}Y_{j,it-1}+X'_{m,it}\beta_m+A_{m,i}}} \frac{1}{1+e^{\sum_{j=1}^{M}\gamma_{mj}k_j+X'_{m,it+1}\beta_m+A_{m,i}}} \prod_{m=1}^{M}(1+e^{\sum_{j=1}^{M}\gamma_{mj}Y_{j,it-1}+X'_{m,it}\beta_m+A_{m,i}})$$

$$= \prod_{m=1}^{M} \frac{e^{k_m(\sum_{j=1}^{M}\gamma_{mj}k_j+X'_{m,it+1}\beta_m+A_{m,i})}}{1+e^{\sum_{j=1}^{M}\gamma_{mj}k_j+X'_{m,it+1}\beta_m+A_{m,i}}}$$

$$= \pi_t^{k|k}(A_i,X_i)$$

which concludes the proof.

M Proof of Lemma 5

By definition, for $T \geq 3$, and for t,s such that $T-1 \geq t > s \geq 1$:

$$\mathbb{E}\left[\zeta_\theta^{k|k}(Y_{it-1}^{t+1},Y_{is-1}^{s},X_i)|Y_{i0},Y_{i1}^{s-1},X_i,A_i\right] = P(Y_{is}=k|Y_{i0},Y_{i1}^{s-1},X_i,A_i)+$$

$$\sum_{l\in\mathcal{Y}\setminus\{k\}} \omega_{t,s,l}^{k|k}(\theta)\mathbb{E}\left[\mathbb{1}\{Y_{is}=l\}\phi_\theta^{k|k}(Y_{it-1}^{t+1},X_i)|Y_{i0},Y_{i1}^{s-1},X_i,A_i\right]$$

$$= \prod_{m=1}^{M} \frac{e^{k_m(\mu_{m,s}(\theta)+A_{m,i})}}{1+e^{\mu_{m,s}(\theta)+A_{m,i}}} + \sum_{l\in\mathcal{Y}\setminus\{k\}} \omega_{t,s,l}^{k|k}(\theta)\pi_t^{k|k}(A_i,X_i)P(Y_{is}=l|Y_{i0},Y_{i1}^{s-1},X_i,A_i)$$

$$= \prod_{m=1}^{M} \frac{e^{k_m(\mu_{m,s}(\theta)+A_{m,i})}}{1+e^{\mu_{m,s}(\theta)+A_{m,i}}} + \sum_{l\in\mathcal{Y}\setminus\{k\}} \left[1-e^{\sum_{j=1}^{M}(l_j-k_j)\left[\kappa_{j,t}^{k|k}(\theta)-\mu_{j,s}(\theta)\right]}\right] \prod_{m=1}^{M} \frac{e^{k_m(\kappa_{m,t}^{k|k}(\theta)+A_{m,i})}}{1+e^{\kappa_{m,t}^{k|k}(\theta)+A_{m,i}}} \frac{e^{l_m(\mu_{m,s}(\theta)+A_{m,i})}}{1+e^{\mu_{m,s}(\theta)+A_{m,i}}}$$

$$= \prod_{m=1}^{M} \frac{e^{k_m(\kappa_{m,t}^{k|k}(\theta)+A_{m,i})}}{1+e^{\kappa_{m,t}^{k|k}(\theta)+A_{m,i}}}$$

$= \pi_t^{k|k}(A_i, X_i)$

The first line follows from the measurability of the weight $\omega_{t,s,l}^{k|k}(\theta)$ with respect to the conditioning set and the linearity of conditional expectations. The second line uses the definition of $\mu_{j,s}(\theta)$ and follows from the law of iterated expectations and Lemma 5. The third line makes use of the definition of $\kappa_{m,t}^{k|k}(\theta)$ and $\omega_{t,s,l}^{k|k}(\theta)$ and the penultime line uses Appendix Lemma 9.

N Dynamic network formation with transitivity

Graham (2013) studies a variant of model (7) to describe network formation amongst groups of 3 individuals. This is a panel data setting where a large sample of many such groups and the evolution of their social ties are observed over $T = 3$ periods (4 counting the initial condition). Interactions are assumed undirected and modelled at the dyad level as:

$$\begin{aligned} D_{ijt} &= \mathbb{1}\left\{\gamma_0 D_{ijt-1} + \delta_0 R_{ijt-1} + A_{ij} - \epsilon_{ijt} \geq 0\right\} \quad t = 1, \ldots, T \\ R_{ijt-1} &= D_{ikt-1} D_{jkt-1} \end{aligned} \tag{12}$$

where i, j, k denote the 3 different agents and $D_{ijt} \in \{0, 1\}$ encodes the presence or absence of a link between agent i and agent j at time t. The network $D_0 \in \{0, 1\}^3$ forms the initial condition. The parameter γ_0 captures state dependence while δ_0 captures transitivity in relationships, i.e the effect of sharing friends in common on the propensity to establish friendships. Finally, A_{ij} is an unrestricted dyad level fixed effect that could potentiall capture unobserved homophily and ϵ_{ijt} is a standard logistic shock, iid over time and individuals. While Graham (2013) establishes identification of (γ_0, δ_0) for $T = 3$ via a conditional likelihood approach in the spirit of Chamberlain (1985), one limitation of the model is the absence of other covariates, in particular time-specific effects. Controlling for such effects can be essential to adequately capture important variation in social dynamics: think about the persistent impact of Covid-19 on all types of social interactions. A relevant extension is thus:

$$
\begin{aligned}
D_{ijt} &= \mathbb{1}\left\{\gamma_0 D_{ijt-1} + \delta_0 D_{ikt-1}D_{jkt-1} + X'_{ijt}\beta_0 + A_{ij} - \epsilon_{ijt} \geq 0\right\} \quad t = 1, \ldots, T \\
R_{ijt-1} &= D_{ikt-1}D_{jkt-1}
\end{aligned}
\tag{13}
$$

Letting $\mathbb{D} = \{0,1\}^3$ denote the support of the network $D_t = (D_{ijt}, D_{ikt}, D_{jkt})$, it is straightforward to see that the results developed for the VAR(1) case can be repurposed to suit model (13) . For $T = 3$, an adaptation of Lemma 4 yields 8 possible transition functions given by:

$$
\phi_\theta^{d|d}(D_3, D_2, D_1, X) = \mathbb{1}\{D_2 = d\} \exp\left(\sum_{i<j}(D_{ij3} - d_{ij2})[\gamma(D_{ij1} - d_{ij2}) - \Delta R_{ij1}\delta - \Delta X'_{ij2}\beta]\right), \quad d \in \mathbb{D}
$$

An adaptation of Lemma 5 implies that we can construct another 8 transition functions given by

$$
\zeta_\theta^{d|d}(D_3, D_2, D_1, D_0, X) = \mathbb{1}\{D_1 = d\} + \sum_{d' \in \mathbb{D}\setminus\{d\}} \omega_{2,1,d'}^{d|d}(\theta)\mathbb{1}\{D_1 = l\}\phi_\theta^{d|d}(D_3, D_2, D_2, X), \quad d \in \mathbb{D}
$$

where

$$
\begin{aligned}
\mu_{ij,1}(\theta) &= \gamma D_{ij0} + \delta R_{ij0} + X'_{ij1}\beta \\
\kappa_{ij,2}^{d|d}(\theta) &= \gamma d_{ij} + \delta r_{ij} + X'_{ij3}\beta \\
\omega_{2,1,d'}^{d|d}(\theta) &= 1 - e^{\sum_{i<j}(d'_{ij} - d_{ij})\left[\kappa_{ij,2}^{d|d}(\theta) - \mu_{ij,1}(\theta)\right]}
\end{aligned}
$$

Therefore, for $T = 3$, 8 moment functions that all meaningfully depend on the model parameter are:

$$
\psi_\theta^{d|d}(D_3, D_2, D_1, D_0, X) = \phi_\theta^{d|d}(D_3, D_2, D_1, X) - \zeta_\theta^{d|d}(D_3, D_2, D_1, D_0, X), \quad d \in \mathbb{D}
$$

Their validity, in the sense of verifying equation (1), follows from the law of iterated expectations.

O Proof of Lemma 6

Let

$$\phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i) = \mathbb{1}\{Y_{it} = k\} e^{\sum_{c \in \mathcal{Y} \setminus \{k\}} \mathbb{1}\{Y_{it+1}=c\} \left(\sum_{j \in \mathcal{Y}} (\gamma_{cj} - \gamma_{kj}) \mathbb{1}(Y_{it-1}=j) + \gamma_{kk} - \gamma_{ck} + \Delta X_{ikt+1}' \beta_k - \Delta X_{ict+1}' \beta_c\right)}$$

We verify the claim by direct computation. We have:

$$\begin{aligned}
&\mathbb{E}\left[\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) | Y_{i0}, Y_{i1}^{t-1}, X_i\right] = P(Y_{it} = k | Y_i^0, Y_{i1}^{t-1}, X_i, A_i) \times \\
&\sum_{l \in \mathcal{Y}} P(Y_{it+1} = l | Y_i^0, Y_{i1}^{t-1}, Y_{it} = k, X_i, A_i) \phi_\theta^{k|k}(l, k, Y_{it-1}, X_i) \\
&= \frac{e^{\sum_{c=0}^C \gamma_{kc} \mathbb{1}(Y_{it-1}=c) + X_{ikt}' \beta_k + A_{ik}}}{\sum_{j=0}^C e^{\sum_{c=0}^C \gamma_{jc} \mathbb{1}(Y_{it-1}=c) + X_{ijt}' \beta_j + A_{ij}}} \times \\
&\sum_{l \in \mathcal{Y}} \frac{e^{\gamma_{lk} + X_{ilt+1}' \beta_l + A_{il}}}{\sum_{j=0}^C e^{\gamma_{jk} + X_{ijt+1}' \beta_j + A_{ij}}} \phi_\theta^{k|k}(l, k, Y_{it-1}, X_i) \\
&= \frac{e^{\sum_{c=0}^C \gamma_{kc} \mathbb{1}(Y_{it-1}=c) + X_{ikt}' \beta_k + A_{ik}}}{\sum_{j=0}^C e^{\sum_{c=0}^C \gamma_{jc} \mathbb{1}(Y_{it-1}=c) + X_{ijt}' \beta_j + A_{ij}}} \times \\
&\left(\frac{e^{\gamma_{kk} + X_{ikt+1}' \beta_k + A_{ik}}}{\sum_{j=0}^C e^{\gamma_{jk} + X_{ijt+1}' \beta_j + A_{ij}}} + \sum_{l \in \mathcal{Y} \setminus \{k\}} \frac{e^{\gamma_{lk} + X_{ilt+1}' \beta_l + A_{il}}}{\sum_{j=0}^C e^{\gamma_{jk} + X_{ijt+1}' \beta_j + A_{ij}}} e^{\left(\sum_{j=0}^C (\gamma_{lj} - \gamma_{kj}) \mathbb{1}(Y_{it-1}=j) + \gamma_{kk} - \gamma_{lk} + \Delta X_{ikt+1}' \beta_k - \Delta X_{ilt+1}' \beta_l\right)} \right) \\
&= \frac{e^{\sum_{c=0}^C \gamma_{kc} \mathbb{1}(Y_{it-1}=c) + X_{ikt}' \beta_k + A_{ik}}}{\sum_{j=0}^C e^{\sum_{c=0}^C \gamma_{jc} \mathbb{1}(Y_{it-1}=c) + X_{ijt}' \beta_j + A_{ij}}} \times \frac{e^{\gamma_{kk} + X_{ikt+1}' \beta_k + A_{ik}}}{\sum_{j=0}^C e^{\gamma_{jk} + X_{ijt+1}' \beta_j + A_{ij}}} \\
&+ \frac{e^{\gamma_{kk} + X_{ikt+1}' \beta_k + A_{ik}}}{\sum_{j=0}^C e^{\sum_{c=0}^C \gamma_{jc} \mathbb{1}(Y_{it-1}=c) + X_{ijt}' \beta_j + A_{ij}}} \times \sum_{l \in \mathcal{Y} \setminus \{k\}} \frac{1}{\sum_{j=0}^C e^{\gamma_{jk} + X_{ijt+1}' \beta_j + A_{ij}}} e^{\sum_{j=0}^C \gamma_{lj} \mathbb{1}(Y_{it-1}=j) + X_{ilt}' \beta_l + A_{il}} \\
&= \frac{e^{\gamma_{kk} + X_{ikt+1}' \beta_k + A_{ik}}}{\sum_{j=0}^C e^{\sum_{c=0}^C \gamma_{jc} \mathbb{1}(Y_{it-1}=c) + X_{ijt}' \beta_j + A_{ij}}} \frac{1}{\sum_{j=0}^C e^{\gamma_{jk} + X_{ijt+1}' \beta_j + A_{ij}}} \sum_{l \in \mathcal{Y}} e^{\sum_{j=0}^C \gamma_{lj} \mathbb{1}(Y_{it-1}=j) + X_{ilt}' \beta_l + A_{il}} \\
&= \frac{e^{\gamma_{kk} + X_{ikt+1}' \beta_k + A_{ik}}}{\sum_{j=0}^C e^{\gamma_{jk} + X_{ijt+1}' \beta_j + A_{ij}}}
\end{aligned}$$

$$= \pi_t^{k|k}(A_i, X_i)$$

which concludes the proof.

P Proof of Lemma 7

By construction for $T \geq 3$, and t, s such that $T-1 \geq t > s \geq 1$,

$$\begin{aligned}
&\mathbb{E}\left[\zeta_{\theta_0}^{0|0}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= P(Y_{is} = 0|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i) \\
&+ \sum_{l \in \mathcal{Y}\setminus\{0\}} \omega_{t,s,l}^{0|0}(\theta)\mathbb{E}\left[\mathbb{1}\{Y_{is} = l\}\mathbb{E}\left[\phi_\theta^{0|0}(Y_{it-1}^{t+1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{1}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} + \sum_{l=1}^{C} \omega_{t,s,l}^{0|0}(\theta)\mathbb{E}\left[\mathbb{1}\{Y_{is} = l\}|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right]\pi_t^{0|0}(A_i, X_i) \\
&= \frac{1}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} + \sum_{l=1}^{C}\left(1 - e^{(\kappa_{l,t}^{0|0}(\theta)-\mu_{l,s}(\theta))}\right)\frac{e^{\mu_{l,s}(\theta)+A_{il}}}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}}\frac{1}{1+\sum_{c=1}^{C} e^{\kappa_{c,t}^{0|0}(\theta)+A_{ic}}} \\
&= \frac{1}{1+\sum_{c=1}^{C} e^{\kappa_{c,t}^{0|0}(\theta)+A_{ic}}} \\
&= \pi_t^{0|0}(A_i, X_i)
\end{aligned}$$

The first line follows from the measurability of the weight $\omega_{t,s,l}^{0|0}(\theta)$ with respect to the conditioning set and the linearity of conditional expectations. The second line uses the definition of $\mu_{c,s}(\theta)$ and follows from the law of iterated expectations and Lemma 6. The third line makes use of the definition of $\kappa_{c,t}^{0|0}(\theta)$, $\omega_{t,s,l}^{0|0}(\theta)$ and the normalization $\gamma_{c0} = \gamma_{0c} = 0, A_{0c} = 0$ for all $c \in \mathcal{Y}$. The penultime line uses Appendix Lemma 8.

Likewise, for all $k \in \mathcal{Y} \setminus \{0\}$,

$$\begin{aligned}
&\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= P(Y_{is} = k|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i) \\
&+ \sum_{l \in \mathcal{Y}\setminus\{k\}} \omega_{t,s,l}^{k|k}(\theta)\mathbb{E}\left[\mathbb{1}\{Y_{is} = l\}\mathbb{E}\left[\phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{\mu_{k,s}(\theta)+A_{ik}}}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} + \sum_{l\in\mathcal{Y}\setminus\{k\}} \omega_{t,s,l}^{k|k}(\theta)\mathbb{E}\left[\mathbb{1}\{Y_{is}=l\}|Y_{i0},Y_{i1}^{s-1},X_i,A_i\right]\pi_t^{k|k}(A_i,X_i) \\
&= \frac{e^{\mu_{k,s}(\theta)+A_{ik}}}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} \\
&+ \sum_{l\in\mathcal{Y}\setminus\{k\}} \left(1-e^{(\kappa_{l,t}^{k|k}(\theta)-\mu_{l,s}(\theta))-(\kappa_{k,t}^{k|k}(\theta)-\mu_{k,s}(\theta))}\right) \frac{e^{\mu_{l,s}(\theta)+A_{il}}}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} \frac{e^{\kappa_{k,t}^{k|k}(\theta)+A_{ik}}}{1+\sum_{c=1}^{C} e^{\kappa_{c,t}^{k|k}(\theta)+A_{ic}}} \\
&= \frac{e^{\mu_{k,s}(\theta)+A_{ik}}}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} + \left(1-e^{-\kappa_{k,t}^{k|k}(\theta)+\mu_{k,s}(\theta)}\right) \frac{1}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} \frac{e^{\kappa_{k,t}^{k|k}(\theta)+A_{ik}}}{1+\sum_{c=1}^{C} e^{\kappa_{c,t}^{k|k}(\theta)+A_{ic}}} \\
&+ \sum_{\substack{l=1\\ l\neq k}}^{C} \left(1-e^{(\kappa_{l,t}^{k|k}(\theta)-\mu_{l,s}(\theta))-(\kappa_{k,t}^{k|k}(\theta)-\mu_{k,s}(\theta))}\right) \frac{e^{\mu_{l,s}(\theta)+A_{il}}}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} \frac{e^{\kappa_{k,t}^{k|k}(\theta)+A_{ik}}}{1+\sum_{c=1}^{C} e^{\kappa_{c,t}^{k|k}(\theta)+A_{ic}}} \\
&= \frac{e^{\kappa_{k,t}^{k|k}(\theta)+A_{ik}}}{1+\sum_{c=1}^{C} e^{\kappa_{c,t}^{k|k}(\theta)+A_{ic}}} \\
&= \pi_t^{k|k}(A_i,X_i)
\end{aligned}
$$

The first line follows from the measurability of the weight $\omega_{t,s,l}^{k|k}(\theta)$ with respect to the conditioning set and the linearity of conditional expectations. The second line uses the definition of $\mu_{k,s}(\theta)$ and follows from the law of iterated expectations and Lemma 6. The third line makes use of the definition of $\kappa_{c,t}^{k|k}(\theta)$ and $\omega_{t,s,l}^{k|k}(\theta)$. The fourth line uses the fact that $\kappa_{0,t}^{k|k}(\theta) = \mu_{0,s}(\theta) = 0$ due to the normalization $\gamma_{c0} = \gamma_{0c} = 0, A_{0c} = 0$ for all $c \in \mathcal{Y}$. The penultime line uses Appendix Lemma 8.

Q Proof of Theorem 2

In what follows, we will drop the cross-sectional subscript i to economize on space.

The proof is based on an application of Theorem 3.2 in Newey (1990). First, in paragraphs A)-B), we verify that the model is smooth: the likelihood is continuous in the common parameter θ with a mean square continuously differentiable square root. We follow a line of argument similar to Davezies et al. (2021) who studied static models. Second, we characterize the nonparametric tangent set $\mathcal{T}$ and its orthogonal complement $\mathcal{T}^{\perp}$ in paragraphs C)-D). Third, we verify in paragraphs E) that the efficient score - the projection of the score onto

$\mathcal{T}^{\perp}$ - coincides with the efficient moment function for $\mathbb{E}\left[\psi_{\theta_0}(Y_0, Y, X)|Y_0, X\right] = 0$, namely

$$\psi_{\theta_0}^{eff}(Y_0, Y, X) = -D(Y_0, X)'\Sigma(Y_0, X)^{-1}\psi_{\theta_0}(Y_0, Y, X) \tag{14}$$

(see e.g Chamberlain (1987))

A) Preliminary calculations

For the AR(1) model, the conditional density of history Y given initial condition Y_0, regressors X and fixed effect A is $f(Y|Y_0, X, A; \theta) = \prod_{t=1}^{T} \frac{e^{Y_t(\gamma Y_{t-1} + X_t'\beta + A)}}{\left(1+e^{\gamma Y_{t-1}+X_t'\beta+A}\right)}$. This implies

$$\begin{aligned} \ln f(Y|Y_0, X, A; \theta) = &\sum_{t=1}^{T} Y_t(\gamma Y_{t-1} + X_t'\beta + A) - \sum_{t=1}^{T} Y_{t-1} \ln\left(1 + e^{\gamma + X_t'\beta + A}\right) \\ &- \sum_{t=1}^{T} (1 - Y_{t-1}) \ln\left(1 + e^{X_t'\beta + A}\right) \end{aligned}$$

which is continously differentiable in θ. Hence

$$\begin{aligned} \frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \gamma} &= \sum_{t=1}^{T} Y_t \left(Y_{t-1} - \frac{e^{\gamma + X_t'\beta + A}}{1 + e^{\gamma + X_t'\beta + A}}\right) \\ \frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \beta} &= \sum_{t=1}^{T} X_t \left(Y_t - Y_{t-1}\frac{e^{\gamma + X_t'\beta + A}}{1 + e^{\gamma + X_t'\beta + A}} - (1 - Y_{t-1})\frac{e^{X_t'\beta + A}}{1 + e^{X_t'\beta + A}}\right) \end{aligned}$$

and because $\mathcal{Y} = \{0, 1\}$, we have

$$\begin{aligned} \left|\frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \gamma}\right| &\leq T \\ \left|\frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \beta}\right| &\leq \sum_{t=1}^{T} |X_t| \end{aligned} \tag{15}$$

B) The model is *smooth*

The conditional likelihood of history Y given regressors X and initial condition Y_0 writes:

$$\mathcal{L}(\theta) = \int f(Y|Y_0, X, A; \theta)\pi(a|Y_0, X)da$$

where $\pi(.|Y_0, X)$ denotes the conditional density of the fixed effect. By (15), the dominated convergence theorem, and Assumption 1 i), the scores for γ and β are given by:

$$S_\gamma = \frac{\partial \ln \mathcal{L}(\theta)}{\partial \gamma} = \mathbb{E}\left[\frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \gamma}|Y_0, Y, X\right]$$
$$S_\beta = \frac{\partial \ln \mathcal{L}(\theta)}{\partial \beta} = \mathbb{E}\left[\frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \beta}|Y_0, Y, X\right]$$

and $\theta \mapsto \mathbb{E}\left[S_\theta S_\theta'\right]$ is continuous. It follows by Lemma 7.6 in Van der Vaart (2000) that $\theta \mapsto \mathcal{L}(\theta)^{\frac{1}{2}}$ is differentiable in quadratic mean, and hence that $\mathcal{L}(\theta)$ is smooth per Definition A.1 in Newey (1990).

C) Nonparametric tangent set

Consider a one-dimensional parametric family for the heterogeneity distribution $\pi(.|Y_0, X; \eta)$ such that $\pi(.|Y_0, X) = \pi(.|Y_0, X; \eta_0)$. Then, the conditional likelihood of the parametric submodel is

$$\mathcal{L}(\theta, \eta) = \int f(Y|Y_0, X, a; \theta)\pi(a|Y_0, X; \eta)da$$

and the score for the nuisance parameter is

$$S_\eta = \frac{\partial \ln \mathcal{L}(\theta, \eta)}{\partial \eta}|_{\eta=\eta_0} = \mathbb{E}\left[\frac{\partial \ln \pi(A|Y_0, X; \eta_0)}{\partial \eta}|Y_0, Y, X\right]$$

Following Hahn (2001), this implies that the nonparametric tangent set is given by

$$\mathcal{T} = \left\{\mathbb{E}[K(A, Y_0, X)|Y_0, Y, X] \text{ such that } \mathbb{E}[K(A, Y_0, X)|Y_0, X] = 0\right\}$$

Having shown that the likelihood is smooth in θ, noting that $\mathcal{T}$ is linear, and that by Assumption 1 ii),

$$\mathbb{E}\left[\psi_{\theta_0}^{eff}(Y_0, Y, X)\psi_{\theta_0}^{eff}(Y_0, Y, X)'\right] = \mathbb{E}\left[D(Y_0, X)'\Sigma(Y_0, X)^{-1}D(Y_0, X)'\right]$$

is nonsingular, all that remains to check are: a) $\psi_{\theta_0}^{eff}(Y_0, Y, X) \in \mathcal{T}^\perp$ and b) $S_\theta - \psi_{\theta_0}^{eff}(Y_0, Y, X) \in \mathcal{T}$. To this end, similarly to Hahn (1997), we shall first show that a) and b) hold conditional on a pair (y_0, x) for the initial condition and the regressors. In other words, we will prove next that $\psi_{\theta_0}^{eff}(y_0, Y, x)$ is the projection of the score onto the orthocomplement of the closed linear space

$$\mathcal{T}_{(y_0,x)} = \left\{\mathbb{E}[K(A,y_0,x)|Y_0 = y_0, Y, X = x] \text{ such that } \mathbb{E}[K(A,y_0,x)|Y_0 = y_0, X = x] = 0\right\}$$

D) Verification of condition a) $\psi^{\mathbf{eff}}_{\theta_0}(\mathbf{y_0},\mathbf{Y},\mathbf{x}) \in \mathcal{T}^{\perp}_{(\mathbf{y_0},\mathbf{x})}$

We begin by characterizing the orthocomplement of the nuisance tangent set $\mathcal{T}^{\perp}_{(y_0,x)}$. By definition, any $g(y_0,Y,x) \in \mathcal{T}^{\perp}_{(y_0,x)}$ is such that for any element of $\mathcal{T}_{(y_0,x)}$, $\mathbb{E}[K(A,y_0,x)|Y_0 = y_0, Y, X = x]$, we have

$$\begin{aligned}
0 &= \mathbb{E}\left[g(y_0,Y,x)\mathbb{E}[K(A,y_0,x)|Y_0 = y_0, Y, X = x]'|Y_0 = y_0, X = x\right] \\
&= \int \mathbb{E}\left[g(y_0,Y,x)|Y_0 = y_0, X = x, A = a\right] K(a,y_0,x)'\pi(a|y_0,x)da
\end{aligned}$$

Since this equality must hold for any K verifying $\mathbb{E}[K(A,y_0,x)|Y_0 = y_0, X = x] = 0$, choosing

$$K(A,y_0,x) = \mathbb{E}\left[g(y_0,Y,x)|Y_0 = y_0, X = x, A\right] - \mathbb{E}\left[g(y_0,Y,x)|Y_0 = y_0, X = x\right]$$

yields

$$0 = \mathbb{V}\Big(\mathbb{E}\left[g(y_0,Y,x)|Y_0 = y_0, X = x, A\right]|Y_0 = y_0, X = x\Big) = 0$$

so that $\mathbb{E}\left[g(y_0,Y,x)|Y_0 = y_0, X = x, A\right] = \mathbb{E}\left[g(y_0,Y,x)|Y_0 = y_0, X = x\right]$. Conversely, it is clear that any g function such that $\mathbb{E}\left[g(y_0,Y,x)|Y_0 = y_0, X = x, A\right]$ is constant in A will be an element of $\mathcal{T}^{\perp}_{(y_0,x)}$ since $\mathbb{E}[K(A,y_0,x)|Y_0 = y_0, X = x] = 0$. We conclude that

$$\begin{aligned}
\mathcal{T}^{\perp}_{(y_0,x)} &= \{g(y_0,Y,x)\,|\,g(y_0,Y,x) = c(y_0,x) + g_*(y_0,Y,x), \quad c(y_0,x) \in \mathbb{R}^{K_x+1}, g_* \in \mathcal{T}^{\perp}_{(y_0,x),*}\} \\
\mathcal{T}^{\perp}_{(y_0,x),*} &= \{g_*(y_0,Y,x)\,|\,\mathbb{E}\left[g_*(y_0,Y,x)|Y_0 = y_0, X = x, A\right] = 0\}
\end{aligned}$$

An important observation is that $\mathcal{T}^{\perp}_{(y_0,x_1),*} = \ker(\mathcal{E}_{y_0,x})^{K_x+1}$, where we recall that the nullspace of the conditional expectation operator $\mathcal{E}_{y_0,x}$ is precisely the set of valid moment functions in the AR(1) model. By Theorem 1, this is a $(2^T - 2T)$-dimensional vector space with basis elements given in Proposition 2. This makes it clear that $\psi^{eff}_{\theta_0}(y_0,Y,x) \in \mathcal{T}^{\perp}_{(y_0,x),*}$ since each of its components is a linear combination of the valid moment functions in Proposition 2 (see specifically equation (14)). Finally since $\mathcal{T}^{\perp}_{(y_0,x),*} \subset \mathcal{T}^{\perp}_{(y_0,x)}$, $\psi^{eff}_{\theta_0}(y_0,Y,x) \in \mathcal{T}^{\perp}_{(y_0,x)}$.

E) Verification of condition b) $\mathbf{S}_{\theta_0} - \psi^{\mathbf{eff}}_{\theta_0}(\mathbf{y_0}, \mathbf{Y}, \mathbf{x}) \in \mathcal{T}_{(\mathbf{y_0},\mathbf{x})}$

We note that since $\mathcal{T}_{(y_0,x)}$ is a closed vector space of a Hilbert space, $\mathcal{T}_{(y_0,x)} = \left(\mathcal{T}^{\perp}_{(y_0,x)}\right)^{\perp}$. Thus, to check condition b) $S_{\theta_0} - \psi^{eff}_{\theta_0}(y_0, Y, x) \in \mathcal{T}_{(y_0,x)}$, we will verify that $\forall g \in \mathcal{T}^{\perp}_{(y_0,x)}$, $\mathbb{E}\left[\left(S_{\theta_0} - \psi^{eff}_{\theta_0}(y_0, Y, x)\right) g(y_0, Y, x)'|Y_0 = y_0, X = x\right] = 0$. Given our previous characterization of $\mathcal{T}^{\perp}_{(y_0,x)}$, it is sufficient to establish that $\mathbb{E}\left[\left(S_{\theta_0} - \psi^{eff}_{\theta_0}(y_0, Y, x)\right) \psi_{\theta_0}(y_0, Y, x)'|Y_0 = y_0, X = x\right] = 0$.

To this end, note that by the Generalized Information Equality (c.f equation (5.1) in Newey and McFadden (1994)) we have

$$\mathbb{E}\left[\frac{\partial \psi_{\theta_0}(y_0, Y, x)}{\partial \theta'}|Y_0 = y_0, X = x\right] = -\mathbb{E}\left[\psi_{\theta_0}(y_0, Y, x)S_{\theta_0}(y_0, Y, x)'|Y_0 = y_0, X = x\right]$$

which implies

$$\begin{aligned}
\psi^{eff}_{\theta_0}(y_0, Y, x) = {} & \mathbb{E}\left[\psi_{\theta_0}(y_0, Y, x)S_{\theta_0}(y_0, Y, x)'|Y_0 = y_0, X = x\right]' \times \\
& \mathbb{E}\left[\psi_{\theta_0}(y_0, Y, x)\psi_{\theta_0}(y_0, Y, x)'|Y_0 = y_0, X = x\right]^{-1} \psi_{\theta_0}(y_0, Y, x) \\
= {} & \mathbb{E}\left[S_{\theta_0}(y_0, Y, x)\psi_{\theta_0}(y_0, Y, x)'|Y_0 = y_0, X = x\right] \times \\
& \mathbb{E}\left[\psi_{\theta_0}(y_0, Y, x)\psi_{\theta_0}(y_0, Y, x)'|Y_0 = y_0, X = x\right]^{-1} \psi_{\theta_0}(y_0, Y, x) \\
= {} & \mathbb{E}^{*}\left[S_{\theta_0}(y_0, Y, x)|\psi_{\theta_0}(y_0, Y, x); Y_0 = y_0, X = x\right]
\end{aligned}$$

where $\mathbb{E}^{*}\left[Z_1|Z_2; W\right]$ denotes the (mean-squared error minimizing) linear predictor of Z_1 on Z_2 given W. Therefore, it immediately follows by properties of conditional linear predictors (e.g Wooldridge (1999), Lemma 4.1) that

$$\mathbb{E}\left[\left(S_{\theta}(y_0, Y, x) - \psi^{eff}_{\theta}(y_0, Y, x)\right) \psi_{\theta}(y_0, Y, x)'|Y_0 = y_0, X = x\right] = 0$$

We conclude that $\psi^{eff}_{\theta_0}(y_0, Y, x)$ is the projection of the score onto $\mathcal{T}^{\perp}_{(y_0,x)}$. It follows that $\psi^{eff}_{\theta_0}(Y_0, Y, X)$ is the projection of the score onto $\mathcal{T}^{\perp}$, i.e it is the efficient score.