Title

Essays in Panel Data and Network Econometrics

Author

Dano, Kevin

Publication Date

2024

Peer reviewed|Thesis/dissertation

Essays in Panel Data and Network Econometrics

by

Kevin Dano

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bryan S. Graham, Chair
Associate Professor Demian Pouzo
Professor James L. Powell

Spring 2024

Essays in Panel Data and Network Econometrics

Abstract

Essays in Panel Data and Network Econometrics

by

Kevin Dano

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Bryan S. Graham, Chair

This dissertation studies how to leverage the unique characteristics of panel and network data, particularly repeated observations and symmetries, to recover the structural parameters of three econometric models of theoretical and applied interest.

In Chapter 1, I study parameter identifiablility and estimation of dynamic discrete choice models with strictly exogenous regressors, fixed effects and logistic errors. Specifications of this kind are popular in Labor Economics and Industrial Organization to disentangle the sources of serial persistence in agents' decisions. The primary challenge lies in the nonlinearity of these models, making the treatment of fixed effects difficult in short panel settings. I introduce a new method that exploits the structure of logit-type probabilities and elementary properties of rational fractions to derive moment restrictions in a broad class of models. This includes binary response models of arbitrary lag order as well as first-order panel vector autoregressions and dynamic multinomial logit models. These moment restrictions are free from the fixed effects and provide a natural way to estimate the common parameters via the Generalized Method of Moments. I further establish the identification of a class of average marginal effects which are often of importance in empirical work. The approach is illustrated through an analysis of the dynamics of drug consumption amongst young people in a nationally representative sample.

In Chapter 2, coauthored with Stéphane Bonhomme and Bryan Graham, we study identification in a binary choice panel data model with a single *predetermined* binary covariate (i.e., a covariate sequentially exogenous conditional on lagged outcomes and covariates). The choice model is indexed by a scalar parameter θ, whereas the distribution of unit-specific heterogeneity, as well as the feedback process that maps lagged outcomes into future covariate realizations, are left unrestricted. This setup departs from Chapter 1 which imposed strict exogeneity of explanatory variables, effectively ruling out any influence of past outcomes on

covariates. In this framework, we provide a simple condition under which θ is never point-identified, no matter the number of time periods available. This condition is satisfied in most models, including the logit one. We also characterize the identified set of θ and show how to compute it using linear programming techniques. While θ is not generally point-identified, its identified set is informative in the examples we analyze numerically, suggesting that meaningful learning about θ may be possible even in short panels with feedback. As a complement, we report calculations of identified sets for an average partial effect, and find informative sets in this case as well.

In Chapter 3, I present an approach to address network endogeneity in a linear social interaction model. I consider a setting wherein individual-specific latent random effects influence both outcomes and link formation modelled as a conditionally independent dyad process. Using the exchangeability properties of the framework, I show that controlling or matching individuals by degree-centrality can be sufficient to eliminate the omitted variable bias induced by endogenous peer selection. I leverage this result and insights from Bramoullé et al. (2009) for the case of exogenous friendships to present two simple strategies for the identification and estimation of social effects. Asymptotic properties of the proposed estimators are derived for clustered samples and I illustrate their performance in Monte Carlo simulations.

A mes parents, N'Dèye Dano et Pierre Dano.

Contents

Acknowledgments

First and foremost, I would like to thank my advisors Bryan Graham, Stéphane Bonhomme, Demian Pouzo and James Powell for their generous support and guidance throughout my graduate studies. Bryan, in particular, deserves special acknowledgment for the time he has dedicated to me, for his encouragements, for being an exceptional mentor, and for his profound influence on my interest in econometrics, panel data and networks. I also want to express my gratitude to Stéphane for his mentorship, for being a source of inspiration, and for his numerous comments and insights that contributed to the improvement of my work. It has been a tremendous privilege to learn from Bryan and Stéphane over the years and to collaborate with them on several projects, including Chapter 2 in this dissertation. I am fortunate to count Demian and Jim as my other two advisors and I am deeply grateful for their constructive feedback and enthusiasm for my research.

I want to express my heartfelt appreciation to my family for their love, care and support. I am eternally grateful to my parents for believing in me and motivating me throughout my academic journey. I am indebted to my mother, N'Dèye Dano, for her passion for my work, for her brilliant suggestions, and for our many discussions of economics that always rekindled my interest in the discipline. I sincerely thank my father, Pierre Dano, for his invaluable teachings and for being an inspiration. I also thank him and my little sister, Julie Dano, for their availability during all these years and for their unwavering encouragements. I also acknowledge the wonderful support of my grandmother Marie Dramé, my aunt Marième Faye, my cousins Yoann Dacruz and Oulimata Diop, my godmother Sylvianne Duvernois and her husband Jean-Pierre Duvernois.

Finally, I extend my gratitude to my friends who made my graduate school experience at UC Berkeley truly memorable. I am especially grateful to Bocar Abdoulaye Ba for his generosity and mentorship, and to Yassine Sbai Sassi for our many fruitful research discussions and exchanges. Special thanks also to Andrea Cerrato, Andrew Tai, Will Sandholtz, Felipe Arteaga, Roberto Hsu Rocha, Nick Gebbia, Nisha Pathak, Leon Lu, Masha Vtorushina, Julien Chancereul, Christine Cai, Andrea Hamaui and Pierre Jaffard for their friendship.

Chapter 1

Transition Probabilities and Moment Restrictions in Dynamic Fixed Effects Logit Models

1.1 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 chapter, 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 (1958a), Chamberlain (1985b) and Magnac (2000) showed that the autoregressive parameter can be consistently estimated by conditional likelihood. This approach

[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.

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 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 chapter, 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 (1985b) 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 observ-

ables 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 $t \in \{p+1, \ldots, T-1\}$ 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 chapter 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 chapter is organized as follows. Section 1.2 presents the setting and our main objective. Section 1.3 introduces some terminology and gives an outline of our procedure to construct moment restrictions. Section 1.4 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 presented for the base case of four waves of data. Section 1.5 discusses extensions to the VAR(1) and the dynamic multinomial logit model with one lag, MAR(1) for short. In Section 1.6, we present an empirical illustration on the dynamics of drug consumption amongst young people and Section 1.7 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 1.8.4. Proofs are gathered in the Appendix.

1.2 Setting, assumptions and objective

Let $i = 1, \ldots, N$ denote a population index and $t = 0, \ldots, T$ 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:

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

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 chapter 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:

$$\mathbb{E}\left[\psi_{\theta_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i, A_i\right] = 0 \tag{1.1}$$

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 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; $\text{Im}(f)$, $\ker(f)$, $\text{rank}(f)$ to denote the image, the nullspace and the rank of a linear map f.

1.3 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 chapter 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 $(y^0, l_1, \dots, l_{t-1})$ 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

[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.

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:

$$\mathbb{E}\left[\phi_{\theta_0}^{k|l_1^t,y_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i, A_i\right] = \pi_t^{k|l_1^t,y_0}(A_i, X_i) \tag{1.2}$$

With these notions in hand, we are ready to describe our two-step approach to derive valid moment functions in the sense of equation (1.1). 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.

1.4 Scalar fixed effect models

1.4.1 Moment restrictions for the AR(1) logit model

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

$$Y_{it} = \mathbb{1}\{\gamma_0 Y_{it-1} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\}, \quad t = 1, \ldots, T \tag{1.3}$$

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 $A_i \in \mathbb{R}$.

1.4.1.1 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

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

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 (1.3) 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 $\text{Im}(\mathcal{E}_{y_0,x})$ 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 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 (1.1), 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 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 1.4.4.1.

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 $t \neq s$, the family $\mathcal{F}_T = \left\{1, (\pi_t^{0|0}(.,x))_{t=0}^{T-1}\right\}$ 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

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

Remark 6 below which characterizes the transition functions associated to each element of $\mathcal{F}_T$.

- In the Cox (1958a) 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.

1.4.1.2 Construction of valid moment functions for the pure model

In the absence of exogenous regressors, model (1.3) simplifies to:

$$Y_{it} = \mathbb{1}\{\gamma_0 Y_{it-1} + A_i - \epsilon_{it} \geq 0\}, \quad t = 1, \ldots, T \tag{1.4}$$

which was first introduced by Cox (1958a) and then revisited in Chamberlain (1985b), 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 (1.1).

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 (1.2), 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:

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

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

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

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

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

[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.

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 (1.4) with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, 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 1.4.3 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 (1.4) 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 $k \in \mathcal{Y}$, $t \in \{2, \ldots, T-1\}$ and $s \in \{1, \ldots, t-1\}$. 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 $\mathbb{E}\left[\psi_\theta(Y_{i1}^3, Y_{i0}^2)|Y_{i0} = 0\right] = 0$, $\psi_\theta(Y_{i1}^3, Y_{i1}^2, 0) = (\psi_\theta^{0|0}(Y_{i1}^3, Y_{i1}^2, 0), \psi_\theta^{1|1}(Y_{i1}^3, Y_{i1}^2, 0))'$.

1.4.1.3 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 (1.3).

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:

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

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 (1.3) with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, let*

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

Then:

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

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

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

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.

[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.

The key, as foreshadowed in Section 1.3 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 AR(1) model with exogenous regressors (the only ones when $T = 3$).

Lemma 3. *In model (1.3) 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 (1.3) 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 (1.1). One particularly relevant set of valid moment functions for reasons explained below is presented in Proposition 2.

Proposition 2. *In model (1.3), 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 (1.9) and (1.10) 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 (1.3) 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

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

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)).

1.4.2 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 $T = 3$. A natural follow-up question is to ask how accurately can θ_0 be estimated in that case, or equivalently what is the semi-parametric information bound. In a corrigendum to Hahn (2001), Gu et al. (2023) confirmed that the conditional likelihood

[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.

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 remain unclear with covariates.

To answer these questions, let $\psi_\theta(Y_{i1}^3, Y_{i0}^1, X_i) = (\psi_\theta^{0|0}(Y_{i1}^3, Y_{i0}^1, X_i), \psi_\theta^{1|1}(Y_{i1}^3, Y_{i0}^1, X_i))'$ where the two components correspond to the valid moment functions of Proposition 2 for $T = 3$. Additionally, let $D(X_i, y_0) = \mathbb{E}\left[\frac{\partial \psi_{\theta_0}(Y_{i1}^3, Y_{i0}^1, X_i)}{\partial \theta'} | Y_{i0} = y_0, X_i\right]$ and let $\Sigma(X_i, y_0) = \mathbb{E}\left[\psi_{\theta_0}(Y_{i1}^3, Y_{i0}^1, X_i)\psi_{\theta_0}(Y_{i1}^3, Y_{i0}^1, X_i)' | Y_{i0} = y_0, X_i\right]$.

Assumption 1. *In model (1.3) with $T = 3$ and initial condition $y_0 \in \{0, 1\}$, the matrix $\mathbb{E}\left[D(X_i, y_0)\Sigma(X_i, y_0)^{-1}D(X_i, y_0)' | Y_{i0} = y_0\right]$ exists and is nonsingular.*

With these notations in hand and under the mild conditions of Assumption 1, Theorem 2 clarifies that the *efficient score* coincides with the efficient moment for the conditional moment problem: $\mathbb{E}\left[\psi_\theta(Y_{i1}^3, Y_{i0}^1, X_i) | Y_{i0} = y_0, X_i\right] = 0$. Put differently, the maximal efficiency with which θ_0 can be estimated is $V_0(y_0) = \mathbb{E}[D(X_i, y_0)\Sigma(X_i, y_0)^{-1}D(X_i, y_0)' | Y_{i0} = y_0]^{-1}$. 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 (1.3) with $T = 3$. Fix an initial condition $y_0 \in \{0, 1\}$ and suppose that Assumption 1 holds. Then, the semiparametric efficiency bound of θ_0 is finite and given by $V_0(y_0) = \mathbb{E}[D(X_i, y_0)\Sigma(X_i, y_0)^{-1}D(X_i, y_0)' | Y_{i0} = y_0]^{-1}$.*

The proof of Theorem 2 only involves careful bookkeeping of some tedious algebra and 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.

1.4.3 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:

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

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

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

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

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 (1.3) 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

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

Appendix Section 1.8.2 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 1.8.2 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.

1.4.4 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

$$Y_{it} = \mathbb{1}\left\{\sum_{r=1}^{p} \gamma_{0r} Y_{it-r} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\right\}, \quad t = 1, \ldots, T \tag{1.5}$$

for known autoregressive order $p > 1$ and vector of initial values
$Y_i^0 = (Y_{i-(p-1)}, \ldots, Y_{i-1}, Y_{i0})' \in \mathcal{Y}^p$, 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:

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

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.

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

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

Then the following result holds:

Theorem 3. *Consider model (1.5) 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 $\mathrm{Im}\left(\mathcal{E}^{(p)}_{y^0,x}\right)$ *and therefore*

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

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

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 $T \leq p+1$, $\mathcal{E}^{(p)}_{y^0,x}$ 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 $T \leq p+1$.

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 1.4). 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 1.8.3.

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. 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

[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.

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 (1.5) with $T \geq p + 1$, for all $t \in \{p, \ldots, T - 1\}$ and $y_1^p \in \mathcal{Y}^p$, let*

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

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

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

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

and

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

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 1.8.3. 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 (1.5), 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.

1.4.5 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 1.8.10.1, 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

$$\begin{aligned}\psi_{\theta,\infty}^{0|0,0}(Y_{i4}, Y_{i3}, Y_{i-1}^2, Z_i) = &-(1 - Y_{i1})(1 - Y_{i2})Y_{i3} \\ &+ \left[e^{X_{i34}'\beta} - 1\right](1 - Y_{i1})(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\ &+ e^{-\gamma_1 Y_{i0} + \gamma_2(1 - Y_{i-1}) + X_{i31}'\beta} Y_{i1}(1 - Y_{i2})(1 - Y_{i3})Y_{i4} \\ &+ e^{-\gamma_1 Y_{i0} - \gamma_2 Y_{i-1} + X_{i41}'\beta} Y_{i1}(1 - Y_{i2})(1 - Y_{i3})(1 - Y_{i4})\end{aligned} \tag{1.6}$$

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 $k \in \{1, \ldots, K_x\}$. 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 (1.5) 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 $\theta_0 = (\gamma_{01}, \ldots, \gamma_{0p}, \beta_0')$ 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 (1.6) 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).

1.4.6 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 1.4.5 for examples of sufficient conditions).

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

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

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 (1.5) 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

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

1.5 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 1.6 where we discuss the empirical application.

1.5.1 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:

$$Y_{m,it} = \mathbb{1}\left\{\sum_{j=1}^{M}\gamma_{0mj}Y_{j,it-1} + X_{m,it}'\beta_{0m} + A_{m,i} - \epsilon_{m,it} \geq 0\right\}, \quad m=1,\ldots,M, \quad t=1,\ldots,T \tag{1.7}$$

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 (1.7) 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:

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

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 (1.7) 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 $(k,l) \in \mathcal{Y}^2$*

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

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,

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

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 (1.7) 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 $k \in \mathcal{Y}$*

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

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

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

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 $\phi_\theta^{k|k}(.)$ and $\zeta_\theta^{k|k}$ 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 $m = 1, \ldots, M$. To illustrate the wider applicability of our two-step method, we show in Appendix 3.2 how one can derive moment restrictions in the dynamic network formation model of Graham (2013) and extensions thereof incorporating exogenous covariates.

1.5.2 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:

$$\pi_t^{k|l}(A_i, X_i) = P(Y_{it+1} = k | Y_{it} = l, X_i, A_i) = \frac{e^{\gamma_{kl} + X'_{ikt+1}\beta_k + A_{ik}}}{\sum_{c=0}^{C} e^{\gamma_{cl} + X'_{ict+1}\beta_j + A_{ic}}}, \quad t = 1, \ldots, T \tag{1.8}$$

with $(k, l) \in \mathcal{Y} = \{0, 1, \ldots, C\}$. 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, $\left(P(Y_{i0} = k | X_i, A_i)\right)_{k=1}^{C}$, 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 $\gamma_{j0} = \gamma_{0j} = 0, A_{0,j} = 0$ for all $j \in \mathcal{Y}$ leaving $\theta = \left((\gamma_{kl})_{k,l \geq 1}, (\beta_l)_{l \geq 0}\right)$ 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 alternatives. In this context, equation (1.8) 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 (1985b) 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:

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

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.

[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 $A_{ik}^* = A_{ik} - A_{i0}$.

Lemma 6. *In model (1.8) with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, let for all $k \in \mathcal{Y}$*

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

Then:

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

Unsurprisingly, given the similarities shared between the MAR(1) and all other specifications discussed in the chapter, 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 (1.8) 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 (1.8) 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 $k \in \mathcal{Y}$*

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

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

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

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.

1.6 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 20 in 2001 to 19 and 24 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 as a valid state of residence, our cross section ultimately consists of $N = 6317$ individuals

[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.

[12]. Following Deza (2015), we then consider the trivariate VAR(1) logit model

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

$m \in \{1, 2, 3\}$ (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 1.5.1. 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

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

[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.

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 1.8.4).

Table 1.1 presents the iterated GMM estimates for the trivariate VAR(1) logit model in 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

[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.

[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.

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.

Table 1.1: Parameter estimates of the trivariate VAR(1) logit

	Iterated GMM			Random Effects			Naive MLE		
	A (I)	M (II)	HD (III)	A (IV)	M (V)	HD (VI)	A (VII)	M (VIII)	HD (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 ≥ 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.*

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. 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.

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

1.7 Conclusion

Dynamic discrete choice models are widely used to study the determinants of repeated decisions made by individuals or firms over time. In this chapter, 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 chapter 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.

[18] A natural idea to gauge the effectiveness of policy interventions would be to compute average marginal effects. However, as mentioned in Section 1.5.1, 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.

1.8 Appendix: proofs, simulations and additional materials

1.8.1 Partial Fraction Decomposition

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

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

□

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

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

□

1.8.2 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:

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

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 $T=3$

$$
\begin{aligned}
\psi^{0|0}_{\theta}(Y^3_1,Y^1_{i0},X_i) &= (1-Y_{i1})\phi^{0|0}_{\theta}(Y^3_{i1},X_i) + e^{(X_{i3}-X_{i1})'\beta-\gamma Y_{i0}}Y_{i1}\phi^{0|0}_{\theta}(Y^3_{i1},X_i) - (1-Y_{i1}) \\
&= e^{(X_{i2}-X_{i3})'\beta}(1-Y_{i1})(1-Y_{i2})Y_{i3} + (1-Y_{i1})(1-Y_{i2})(1-Y_{i3}) \\
&+ e^{(X_{i2}-X_{i1})'\beta+\gamma(1-Y_{i0})}Y_{i1}(1-Y_{i2})Y_{i3} \\
&+ e^{(X_{i3}-X_{i1})'\beta-\gamma Y_{i0}}Y_{i1}(1-Y_{i2})(1-Y_{i3}) \\
&- (1-Y_{i1}) \\
&= (e^{(X_{i2}-X_{i3})'\beta}-1)(1-Y_{i1})(1-Y_{i2})Y_{i3} \\
&+ e^{(X_{i2}-X_{i1})'\beta+\gamma(1-Y_{i0})}Y_{i1}(1-Y_{i2})Y_{i3} \\
&+ e^{(X_{i3}-X_{i1})'\beta-\gamma Y_{i0}}Y_{i1}(1-Y_{i2})(1-Y_{i3}) \\
&- (1-Y_{i1})Y_{i2}
\end{aligned} \tag{1.9}
$$

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 (1.9) corresponds to their moment function m^b_0 which they express in an extensive form. For $Y_{i0}=1$, we get instead m^b_1.

Similarly,

$$\begin{aligned}\psi_\theta^{1|1}(Y_{i1}^3, Y_{i0}^1, X_i) &= Y_{i1}\phi_\theta^{1|1}(Y_{i1}^3, X_i) + e^{-\gamma(1-Y_{i0})-(X_{i3}-X_{i1})'\beta)}(1-Y_{i1})\phi_\theta^{1|1}(Y_{i1}^3, X_i) - Y_{i1} \\ &= e^{(X_{i3}-X_{i2})'\beta}Y_{i1}Y_{i2}(1-Y_{i3}) + Y_{i1}Y_{i2}Y_{i3} \\ &+ e^{(X_{i1}-X_{i2})'\beta+\gamma Y_{i0}}(1-Y_{i1})Y_{i2}(1-Y_{i3}) \\ &+ e^{(X_{i1}-X_{i3})'\beta-\gamma(1-Y_{i0})}(1-Y_{i1})Y_{i2}Y_{i3} \\ &- Y_{i1} \\ &= (e^{(X_{i3}-X_{i2})'\beta}-1)Y_{i1}Y_{i2}(1-Y_{i3}) \\ &+ e^{(X_{i1}-X_{i2})'\beta+\gamma Y_{i0}}(1-Y_{i1})Y_{i2}(1-Y_{i3}) \\ &+ e^{(X_{i1}-X_{i3})'\beta-\gamma(1-Y_{i0})}(1-Y_{i1})Y_{i2}Y_{i3} \\ &- Y_{i1}(1-Y_{i2})\end{aligned} \tag{1.10}$$

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 (1.10) 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 $T = 3$.

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

As indicated in Section 1.4.4.2 , **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 (1.5) 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*

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

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 (1.5) 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*

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

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 (1.5)*
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

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

Then,

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

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.

1.8.4 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.

1.8.4.1 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.

Table 1.2: Performance of GMM estimators for the AR(3)

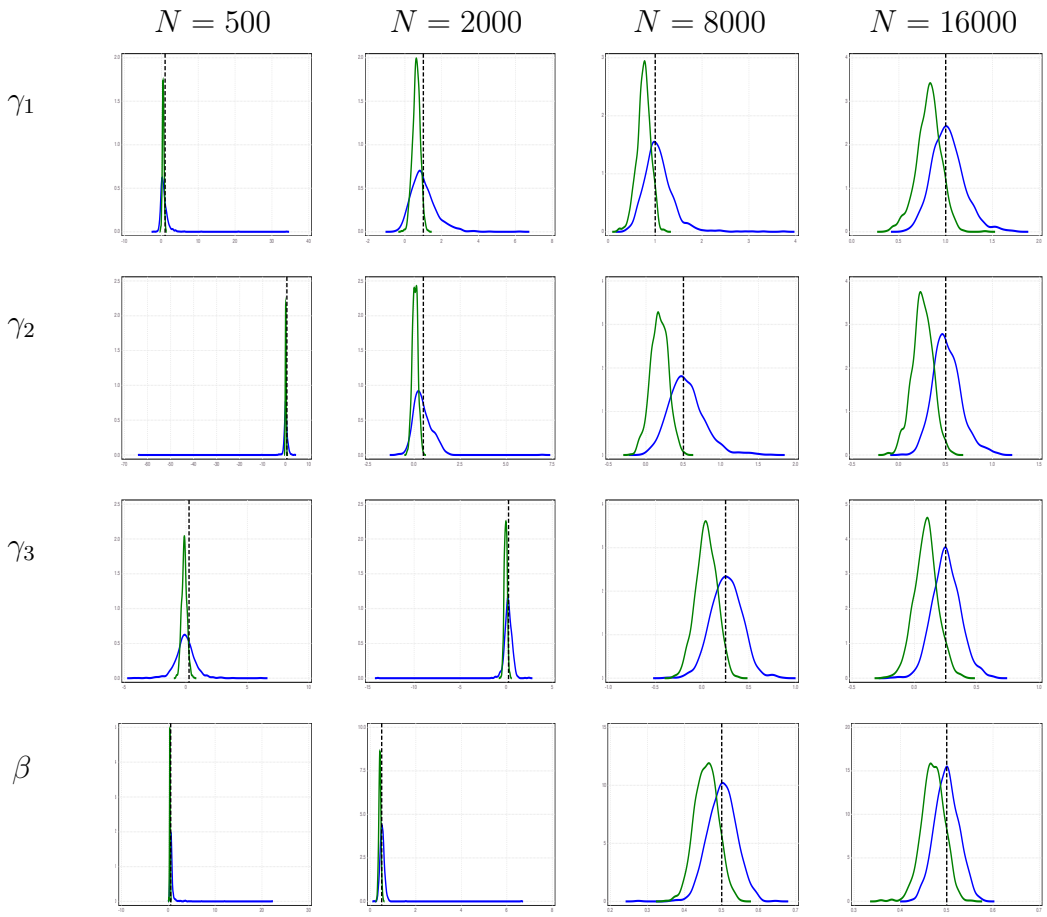
		$\hat{\gamma_1}^a$	$\hat{\gamma_1}^b$	$\hat{\gamma_2}^a$	$\hat{\gamma_2}^b$	$\hat{\gamma_3}^a$	$\hat{\gamma_3}^b$	$\hat{\beta}^a$	$\hat{\beta}^b$
$N = 500$									
	Bias	-0.52	-0.50	-0.51	-0.50	-0.39	-0.32	-0.15	0.10
	MAE	0.52	0.69	0.51	0.58	0.39	0.51	0.15	0.14
$N = 2000$									
	Bias	-0.37	-0.10	-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 1.2 presents the median bias and median absolute errors of the two GMM estimators for each design $N \in \{500, 2000, 8000, 16000\}$. Figure 1.1 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 1.2 and from the fact that all green densities are to the left of the true parameters in Figure 1.1. 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 1.1: 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

[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 $\hat{\theta}^b$.

Figure 1.1: 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.

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 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.

1.8.4.2 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:

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

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 1.6. The results of the simulations are summarized in Table 1.3 and Table 1.4.

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 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 1.6. 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]

[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 1.3: 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* $N = 8000, 16000$.

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 concentrating on the iterated GMM estimator in our empirical application.

Table 1.4: 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* $N = 8000, 16000$.

1.8.5 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:

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

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

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

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

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

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

  and if $y_0 = 1$

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

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

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

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

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

This is the desired result. The terms $\left\{\left(\pi_{t-1}^{y_1|y_1^{t-1}, y_0, \dots, y_{-(p-t)}}(., x)\right)\right)_{y_1^{t-1} \in \mathcal{Y}^{t-1}}\Bigg\}_{t=2}^{p}$ 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 $\dim\left(\ker(\mathcal{E}_{y^0,x}^{(p)})\right) = 2^T - (T-p+1)2^p$

1.8.6 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.

1.8.7 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

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

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:

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

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

1.8.8 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 $T - 1 \geq t > s \geq 1$:

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

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,

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

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.

1.8.9 Proof of Theorem 4

We start by proving the following Lemma

Lemma 11. *In model (1.5), with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, let*

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

Then,

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

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:

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

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

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

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:

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

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

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

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 $k \in \{1, \ldots, p-2\}$, 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 $y_1 = 0, y_{k+1} = 1$

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

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

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

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

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

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

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

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 argu-

ment 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.

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

1.8.10.1 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:

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

Moreover, an application of Lemma 11 gives

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

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

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

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

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

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

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

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

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

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.

1.8.10.2 Proof of Theorem 5

We recall from the discussion of Section 1.4.5 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:

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

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

$$\begin{aligned}\psi^{0|0,0}_{\theta,\infty}(Y_{i4}, Y_{i3}, Y^2_{i-1}, Z_i) = &-(1-Y_{i1})(1-Y_{i2})Y_{i3} \\ &+ \left[e^{X'_{i34}\beta} - 1\right](1-Y_{i1})(1-Y_{i2})(1-Y_{i3})Y_{i4} \\ &+ e^{-\gamma_1 Y_{i0} + \gamma_2(1-Y_{i-1}) + X'_{i31}\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3})Y_{i4} \\ &+ e^{-\gamma_1 Y_{i0} - \gamma_2 Y_{i-1} + X'_{i41}\beta} Y_{i1}(1-Y_{i2})(1-Y_{i3})(1-Y_{i4})\end{aligned} \tag{1.11}$$

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

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

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

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

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

$$\begin{aligned}&\lim_{w_2 \to \infty} \mathbb{E}\left[(1-Y_{i1})Y_{i2} | Y^0_i = y^0, X_i \in \mathcal{X}_s, W_{i2} = w_2\right] \\ &= \lim_{w_2 \to \infty} \int \frac{e^{\gamma_{02} y_0 + x'_2\beta_0 + a}}{1 + e^{\gamma_{02} y_0 + x'_2\beta_0 + a}} \frac{1}{1 + e^{\gamma_{01} y_0 + \gamma_{02} y_{i-1} + x'_1\beta_0 + a}} p(a, z | y_0, \mathcal{X}_s, w_2) da dz\end{aligned}$$

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

$$\begin{aligned}&\lim_{w_2 \to \infty} \mathbb{E}\left[(1-Y_{i1})Y_{i2} | Y^0_i = y^0, X_i \in \mathcal{X}_s, W_{i2} = w_2\right] \\ &= \int \lim_{w_2 \to \infty} \frac{e^{\gamma_{02} y_0 + x'_2\beta_0 + a}}{1 + e^{\gamma_{02} y_0 + x'_2\beta_0 + a}} \frac{1}{1 + e^{\gamma_{01} y_0 + \gamma_{02} y_{i-1} + x'_1\beta_0 + a}} p(a, z | y_0, \mathcal{X}_s, w_2) da dz \\ &= \int 0 \times q(a, z | y_0, \mathcal{X}_s) da dz \\ &= 0\end{aligned}$$

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 (1.6) 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:

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

Hence, by Leibniz integral rule, we get

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

Similarly,

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

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.

1.8.11 Proof of Proposition 3

We recall that by definition,

$$
\begin{aligned}
&\Pi^{k_1^s | l_1^p}_t(y^0, x_1^{t+s}) = \\
&\mathbb{E}\left[P(Y_{it+s} = k_s, \ldots, Y_{it+1} = k_1 \,|\, Y_{it} = l_1, \ldots, Y_{it-(p-1)} = l_p, X_{i1}^{t+s} = x_1^{t+s}, A_i)\right. \\
&\left.\,|\, Y^0_i = y^0, X_{i1}^{t+s} = x_1^{t+s}\right]
\end{aligned}
$$

We have

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

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.

Case 1: $s < p$
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 $t \in \{p, \ldots, T-2\}$, $s \in \{1, \ldots, T-1-t\}$ and $y, \tilde{y} \in \mathcal{Y}^p$, $\gamma_0' y + x_t'\beta_0 \neq \gamma_0'\tilde{y} + x'_{t+s}\beta_0$, $D^{k_1^s|l_1^p}(e^a)$ 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 $(\mu, \lambda_0, \lambda_1, \ldots, \lambda_{s-1}) \in \mathbb{R}^{s+1}$ such that:

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

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

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

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

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

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.

Case 2: $s \geq p$
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.

1.8.12 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.

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

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

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

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

which concludes the proof.

1.8.13 Proof of Lemma 5

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

$$\begin{aligned}
&\mathbb{E}\left[\zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] = P(Y_{is}=k|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i) \\
&+ \sum_{l\in\mathcal{Y}\setminus\{k\}} \omega_{t,s,l}^{k|k}(\theta)\mathbb{E}\left[\mathbb{1}\{Y_{is}=l\}\phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \prod_{m=1}^{M} \frac{e^{k_m(\mu_{m,s}(\theta)+A_{m,i})}}{1+e^{\mu_{m,s}(\theta)+A_{m,i}}} + \sum_{l\in\mathcal{Y}\setminus\{k\}} \omega_{t,s,l}^{k|k}(\theta)\pi_t^{k|k}(A_i, X_i)P(Y_{is}=l|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i) \\
&= \prod_{m=1}^{M} \frac{e^{k_m(\mu_{m,s}(\theta)+A_{m,i})}}{1+e^{\mu_{m,s}(\theta)+A_{m,i}}} \\
&+ \sum_{l\in\mathcal{Y}\setminus\{k\}} \left[1 - e^{\sum_{j=1}^{M}(l_j-k_j)\left[\kappa_{j,t}^{k|k}(\theta)-\mu_{j,s}(\theta)\right]}\right] \prod_{m=1}^{M} \frac{e^{k_m(\kappa_{m,t}^{k|k}(\theta)+A_{m,i})}}{1+e^{\kappa_{m,t}^{k|k}(\theta)+A_{m,i}}} \frac{e^{l_m(\mu_{m,s}(\theta)+A_{m,i})}}{1+e^{\mu_{m,s}(\theta)+A_{m,i}}}
\end{aligned}$$

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

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.

1.8.14 Dynamic network formation with transitivity

Graham (2013) studies a variant of model (1.7) 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:

$$\begin{aligned} D_{ijt} &= \mathbb{1}\left\{\gamma_0 D_{ijt-1} + \delta_0 R_{ijt-1} + A_{ij} - \epsilon_{ijt} \geq 0\right\} \quad t = 1, \ldots, T \\ R_{ijt-1} &= D_{ikt-1} D_{jkt-1} \end{aligned} \tag{1.12}$$

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 (1985b), 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:

$$\begin{aligned} D_{ijt} &= \mathbb{1}\left\{\gamma_0 D_{ijt-1} + \delta_0 D_{ikt-1} D_{jkt-1} + X_{ijt}'\beta_0 + A_{ij} - \epsilon_{ijt} \geq 0\right\} \quad t = 1, \ldots, T \\ R_{ijt-1} &= D_{ikt-1} D_{jkt-1} \end{aligned} \tag{1.13}$$

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 (1.13) . For $T = 3$, an adaptation of Lemma 4 yields 8 possible transition functions given by:

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

for all $d \in \mathbb{D}$. An adaptation of Lemma 5 implies that we can construct another 8 transition functions given by

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

for all $d \in \mathbb{D}$ 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 (1.1), follows from the law of iterated expectations.

1.8.15 Proof of Lemma 6

Let

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

We verify the claim by direct computation. We have:

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

$$
\left(\frac{e^{\gamma_{kk}+X'_{ikt+1}\beta_k+A_{ik}}}{\sum_{j=0}^{C} e^{\gamma_{jk}+X'_{ijt+1}\beta_j+A_{ij}}} \right.
$$

$$
\left. + \sum_{l\in\mathcal{Y}\setminus\{k\}} \frac{e^{\gamma_{lk}+X'_{ilt+1}\beta_l+A_{il}}}{\sum_{j=0}^{C} e^{\gamma_{jk}+X'_{ijt+1}\beta_j+A_{ij}}} e^{\left(\sum_{j=0}^{C}(\gamma_{lj}-\gamma_{kj})\mathbb{1}(Y_{it-1}=j)+\gamma_{kk}-\gamma_{lk}+\Delta X'_{ikt+1}\beta_k-\Delta X'_{ilt+1}\beta_l\right)} \right)
$$

$$
= \frac{e^{\sum_{c=0}^{C}\gamma_{kc}\mathbb{1}(Y_{it-1}=c)+X'_{ikt}\beta_k+A_{ik}}}{\sum_{j=0}^{C} e^{\sum_{c=0}^{C}\gamma_{jc}\mathbb{1}(Y_{it-1}=c)+X'_{ijt}\beta_j+A_{ij}}} \times \frac{e^{\gamma_{kk}+X'_{ikt+1}\beta_k+A_{ik}}}{\sum_{j=0}^{C} e^{\gamma_{jk}+X'_{ijt+1}\beta_j+A_{ij}}}
$$

$$
+ \frac{e^{\gamma_{kk}+X'_{ikt+1}\beta_k+A_{ik}}}{\sum_{j=0}^{C} e^{\sum_{c=0}^{C}\gamma_{jc}\mathbb{1}(Y_{it-1}=c)+X'_{ijt}\beta_j+A_{ij}}} \times \sum_{l\in\mathcal{Y}\setminus\{k\}} \frac{1}{\sum_{j=0}^{C} e^{\gamma_{jk}+X'_{ijt+1}\beta_j+A_{ij}}} e^{\sum_{j=0}^{C}\gamma_{lj}\mathbb{1}(Y_{it-1}=j)+X'_{ilt}\beta_l+A_{il}}
$$

$$
= \frac{e^{\gamma_{kk}+X'_{ikt+1}\beta_k+A_{ik}}}{\sum_{j=0}^{C} e^{\sum_{c=0}^{C}\gamma_{jc}\mathbb{1}(Y_{it-1}=c)+X'_{ijt}\beta_j+A_{ij}}} \frac{1}{\sum_{j=0}^{C} e^{\gamma_{jk}+X'_{ijt+1}\beta_j+A_{ij}}} \sum_{l\in\mathcal{Y}} e^{\sum_{j=0}^{C}\gamma_{lj}\mathbb{1}(Y_{it-1}=j)+X'_{ilt}\beta_l+A_{il}}
$$

$$
= \frac{e^{\gamma_{kk}+X'_{ikt+1}\beta_k+A_{ik}}}{\sum_{j=0}^{C} e^{\gamma_{jk}+X'_{ijt+1}\beta_j+A_{ij}}}
$$

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

which concludes the proof.

1.8.16 Proof of Lemma 7

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

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

$$
= P(Y_{is}=0|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i)
$$

$$
+ \sum_{l\in\mathcal{Y}\setminus\{0\}} \omega_{t,s,l}^{0|0}(\theta)\mathbb{E}\left[\mathbb{1}\{Y_{is}=l\}\mathbb{E}\left[\phi_\theta^{0|0}(Y_{it-1}^{t+1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right]
$$

$$
= \frac{1}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} + \sum_{l=1}^{C} \omega_{t,s,l}^{0|0}(\theta)\mathbb{E}\left[\mathbb{1}\{Y_{is}=l\}|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right]\pi_t^{0|0}(A_i, X_i)
$$

$$
\begin{aligned}
&= \frac{1}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} + \sum_{l=1}^{C} \left(1 - e^{(\kappa_{l,t}^{0|0}(\theta) - \mu_{l,s}(\theta))}\right) \frac{e^{\mu_{l,s}(\theta)+A_{il}}}{1+\sum_{c=1}^{C} e^{\mu_{c,s}(\theta)+A_{ic}}} \frac{1}{1+\sum_{c=1}^{C} e^{\kappa_{c,t}^{0|0}(\theta)+A_{ic}}} \\
&= \frac{1}{1+\sum_{c=1}^{C} e^{\kappa_{c,t}^{0|0}(\theta)+A_{ic}}} \\
&= \pi_t^{0|0}(A_i, X_i)
\end{aligned}
$$

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 $k \in \mathcal{Y} \setminus \{0\}$,

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

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.

1.8.17 Proof of Theorem 2

In what follows, we will drop the cross-sectional subscript i to economize on space. To avoid excessive repetition, we will detail the argument for the initial condition $Y_0 = 0$. A set of completely symmetric arguments will deliver the result for $Y_0 = 1$ and can be provided upon request. For conciseness, we will further omit the conditioning on the initial condition $Y_0 = 0$ in conditional expectations.

A) Preliminary calculations

The conditional density of history (Y_1, Y_2, Y_3) of the AR(1) model given initial condition Y_0, regressors X and fixed effect A is $f(Y_1, Y_2, Y_3|Y_0, X, A; \theta) = \prod_{t=1}^{3} \frac{e^{Y_t(\gamma Y_{t-1} + X_t'\beta + A)}}{\left(1+e^{\gamma Y_{t-1} + X_t'\beta + A}\right)}$. This implies

$$\begin{aligned}\ln f(Y_1, Y_2, Y_3|Y_0, X, A; \theta) = &\sum_{t=1}^{3} Y_t(\gamma Y_{t-1} + X_t'\beta + A) - \sum_{t=1}^{3} Y_{t-1} \ln\left(1 + e^{\gamma + X_t'\beta + A}\right) \\ &- \sum_{t=1}^{3} (1 - Y_{t-1}) \ln\left(1 + e^{X_t'\beta + A}\right)\end{aligned}$$

and hence

$$\begin{aligned}\frac{\partial \ln f(Y_1, Y_2, Y_3|Y_0, X, A; \theta)}{\partial \gamma} &= \sum_{t=1}^{3} Y_t \left(Y_{t-1} - \frac{e^{\gamma + X_t'\beta + A}}{1 + e^{\gamma + X_t'\beta + A}}\right) \\ \frac{\partial \ln f(Y_1, Y_2, Y_3|Y_0, X, A; \theta)}{\partial \beta} &= \sum_{t=1}^{3} X_t \left(Y_t - Y_{t-1}\frac{e^{\gamma + X_t'\beta + A}}{1 + e^{\gamma + X_t'\beta + A}} - (1 - Y_{t-1})\frac{e^{X_t'\beta + A}}{1 + e^{X_t'\beta + A}}\right)\end{aligned}$$

Our candidate for the efficient score is the efficient moment based on the conditional moment restriction: $\mathbb{E}\left[\psi_\theta(Y_1^3, Y_0^1, X)|Y_0 = 0, X\right] = 0$. By Chamberlain (1987), it is given by,

$$\psi_\theta^{eff}(Y_1^3, X) = -\Omega(X)\psi_\theta(Y_1^3, Y_0^1, X)$$

where $\Omega(X) = D(X)'\Sigma(X)^{-1}$ (recall that we are omitting the dependence on the initial condition $Y_0 = 0$ here). The following expressions for $D(X), \Sigma(X), \Omega(X)$ are useful for the derivations ahead:

$$D_{11}(X) = e^{X_{21}'\beta + \gamma} P_{101}(X)$$

$$
\begin{aligned}
D_{21}(X) &= -e^{X_{13}'\beta-\gamma}P_{011}(X) \\
D_{1j}(X) &= X_{23,j-1}e^{X_{23}'\beta}P_{001}(X) + X_{21,j-1}e^{X_{21}'\beta+\gamma}P_{101}(X) + X_{31,j-1}e^{X_{31}'\beta}P_{100}(X), \\
&\quad j = 2,\ldots,K+1 \\
D_{2j}(X) &= X_{32,j-1}e^{X_{32}'\beta}P_{110}(X) + X_{12,j-1}e^{X_{12}'\beta}P_{010}(X) + X_{13,j-1}e^{X_{13}'\beta-\gamma}P_{011}(X), \\
&\quad j = 2,\ldots,K+1 \\
\Sigma_{11}(X) &= (e^{X_{23}'\beta}-1)^2P_{001}(X) + e^{2X_{21}'\beta+2\gamma}P_{101}(X) + e^{2X_{31}'\beta}P_{100}(X) + P_{01}(X) \\
\Sigma_{22}(X) &= (e^{X_{32}'\beta}-1)^2P_{110}(X) + e^{2X_{12}'\beta}P_{010}(X) + e^{2X_{13}'\beta-2\gamma}P_{011}(X) + P_{10}(X) \\
\Sigma_{12}(X) &= \Sigma_{21}(X) \\
&= -\left(e^{X_{21}'\beta+\gamma}P_{101}(X) + e^{X_{31}'\beta}P_{100}(X) + e^{X_{12}'\beta}P_{010}(X) + e^{X_{13}'\beta-\gamma}P_{011}(X)\right) \\
det\left(\Sigma(X)\right) &= \Sigma_{11}(X)\Sigma_{22}(X) - \Sigma_{12}(X)^2 \\
\Omega_{j1}(X) &= \frac{1}{det\left(\Sigma(X)\right)}\left(D_{1j}(X)\Sigma_{22}(X) - D_{2j}(X)\Sigma_{12}(X)\right), \quad j = 1,\ldots,K+1 \\
\Omega_{j2}(X) &= \frac{1}{det\left(\Sigma(X)\right)}\left(-D_{1j}(X)\Sigma_{12}(X) + D_{2j}(X)\Sigma_{11}(X)\right), \quad j = 1,\ldots,K+1
\end{aligned}
$$

were I use the shorthand $P_{y_1\ldots y_n}(X) = P(Y_1 = y_1,\ldots,Y_n = y_n|Y_0 = 0, X)$

B) Scores and nonparametric tangent set

With $T = 3$, the conditional likelihood of history (Y_1, Y_2, Y_3) given $X = x, Y_0 = y_0$ writes:

$$\mathcal{L}(\theta) = \int f(Y_1, Y_2, Y_3|y_0, x, a; \theta)\pi(a|y_0, x)da$$

where $\pi(.|y_0, x)$ denotes the conditional density of A given $X = x, Y_0 = y_0$. Consider a scalar parametric submodel 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

$$\mathcal{L}(\theta, \eta) = \int f(Y_1, Y_2, Y_3|y_0, x, a; \theta)\pi(a|y_0, x; \eta)da$$

Define

$$
\begin{aligned}
C_{y_1y_2y_3}(x_t) &= \mathbb{E}\left[\frac{e^{\gamma+x_t'\beta+A}}{1+e^{\gamma+x_t'\beta+A}}|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, X = x\right] \\
B_{y_1y_2y_3}(x_t) &= \mathbb{E}\left[\frac{e^{x_t'\beta+A}}{1+e^{x_t'\beta+A}}|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, X = x\right]
\end{aligned}
$$

Careful bookkeeping yield the following scores for γ and β

$$\begin{aligned} S_\gamma &= \frac{\partial \ln \mathcal{L}(\theta,\eta)}{\partial \gamma} = \mathbb{E}\left[\frac{\partial \ln f(Y_1,Y_2,Y_3|Y_0,X,A;\theta)}{\partial \gamma}|Y_1,Y_2,Y_3,X=x\right] \\ &= \big(1-C_{111}(x_2)+1-C_{111}(x_3)\big)\,Y_1Y_2Y_3 + (1-C_{110}(x_2)-C_{110}(x_3))Y_1Y_2(1-Y_3) \\ &- C_{101}(x_2)Y_1(1-Y_2)Y_3 - C_{100}(x_2)Y_1(1-Y_2)(1-Y_3) \\ &+ (1-C_{011}(x_3))(1-Y_1)Y_2Y_3 - C_{010}(x_3)(1-Y_1)Y_2(1-Y_3) \end{aligned} \tag{1.14}$$

and

$$\begin{aligned} S_\beta &= \frac{\partial \ln \mathcal{L}(\theta,\eta)}{\partial \beta} = \mathbb{E}\left[\frac{\partial \ln f(Y_1,Y_2,Y_3|Y_0,X,A;\theta)}{\partial \beta}|Y_1,Y_2,Y_3,X=x\right] \\ &= \big(x_1(1-B_{111}(x_1))+x_2(1-C_{111}(x_2))+x_3(1-C_{111}(x_3))\big)\,Y_1Y_2Y_3 \\ &+ \big(x_1(1-B_{110}(x_1))+x_2(1-C_{110}(x_2))-x_3C_{110}(x_3)\big)\,Y_1Y_2(1-Y_3) \\ &+ \big(x_1(1-B_{101}(x_1))-x_2C_{101}(x_2)+x_3(1-B_{101}(x_3))\big)\,Y_1(1-Y_2)Y_3 \\ &+ \big(x_1(1-B_{100}(x_1))-x_2C_{100}(x_2)-x_3B_{100}(x_3)\big)\,Y_1(1-Y_2)(1-Y_3) \\ &+ \big(-x_1B_{011}(x_1)+x_2(1-B_{011}(x_2))+x_3(1-C_{011}(x_3))\big)\,(1-Y_1)Y_2Y_3 \\ &+ \big(-x_1B_{010}(x_1)+x_2(1-B_{010}(x_2))-x_3C_{010}(x_3)\big)\,(1-Y_1)Y_2(1-Y_3) \\ &+ \big(-x_1B_{001}(x_1)-x_2B_{001}(x_2)+x_3(1-B_{001}(x_3))\big)\,(1-Y_1)(1-Y_2)Y_3 \\ &+ \big(-x_1B_{000}(x_1)-x_2B_{000}(x_2)-x_3B_{000}(x_3)\big)\,(1-Y_1)(1-Y_2)(1-Y_3) \end{aligned}$$

The score for the nuisance parameter is

$$S_\eta = \frac{\partial \ln \mathcal{L}(\theta,\eta_0)}{\partial \eta} = \mathbb{E}\left[\frac{\partial \ln \pi(A|y_0,x;\eta_0)}{\partial \eta}|Y_1,Y_2,Y_3,X=x\right]$$

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

$$\mathcal{T} = \big\{\mathbb{E}[K(A,x)|Y_1,Y_2,Y_3,x] \text{ such that } \mathbb{E}[K(A,x)|x]=0\big\}$$

To prove that ψ_θ^{eff} is semiparametrically efficient, we will verify the conditions for an application of Theorem 3.2 in Newey (1990). Noting that $\mathcal{L}(\theta,\eta)$ is differentiable in θ, that $\mathcal{T}$ is linear, and that by Assumption 1, $\mathbb{E}\left[\psi_\theta^{eff}(Y_1^3,X)\psi_\theta^{eff}(Y_1^3,X)'\right] = \mathbb{E}\left[D(X)\Sigma(X)^{-1}D(X)'\right]$ is non singular, all that remains to check are: i) $\psi_\theta^{eff}(Y_1^3,X)\in\mathcal{T}^\perp$ and ii) $S_\theta - \psi_\theta^{eff}(Y_1^3,X)\in\mathcal{T}$.

C) Verification of condition i) $\psi_\theta^{\mathbf{eff}}(\mathbf{Y_1^3},\mathbf{X})\in\mathcal{T}^\perp$

To verify condition i), let us characterize the orthocomplement of $\mathcal{T}$ which will also be useful to verify condition ii). By definition, any $g(Y_1,Y_2,Y_3,x)\in\mathcal{T}^\perp$ is such that for any element of $\mathcal{T}$, $\mathbb{E}[K(A,x)|Y_1,Y_2,Y_3,x]$, we have

$$0 = \mathbb{E}\left[g(Y_1,Y_2,Y_3,x)\mathbb{E}[K(A,x)|Y_1,Y_2,Y_3,x]|x\right]$$

$$= \int K(a,x)\mathbb{E}\left[g(Y_1,Y_2,Y_3,x)|x,a\right]\pi(a|x)da$$

because this equality must be valid for any $K(a,x)$ verifying $\mathbb{E}[K(A,x)|x]=0$, it must be the case that $\mathbb{V}\left(\mathbb{E}\left[g(Y_1,Y_2,Y_3,x)|x,A\right]|x\right)=0$ or equivalently that $\mathbb{E}\left[g(Y_1,Y_2,Y_3,x)|x,A\right]=\mathbb{E}\left[g(Y_1,Y_2,Y_3,x)|x\right]$. Conversely, this short calculation makes it clear that any g function such that $E\left[g(Y_1,Y_2,Y_3,x)|x,A\right]$ is constant will be an element of $\mathcal{T}^{\perp}$. We conclude that,

$$\mathcal{T}^{\perp} = \{g(Y_1,Y_2,Y_3,x)\,|\,\mathbb{E}\left[g(Y_1,Y_2,Y_3,x)-\mathbb{E}\left[g(Y_1,Y_2,Y_3,x)|x\right]|x,A\right]=0\} = \mathbb{R}+\mathcal{T}_*^{\perp}$$
$$\mathcal{T}_*^{\perp} = \{g_*(Y_1,Y_2,Y_3,x)\,|\,\mathbb{E}\left[g_*(Y_1,Y_2,Y_3,x)|x,A\right]=0\}$$

At this stage, an important observation is that $\mathcal{T}_*^{\perp}$ coincides with the set of valid moment functions in the AR(1) model with $T=3$. By Theorem 1, this is a 2-dimensional space when $T=3$ with basis elements $\psi_\theta^{0|0}(Y_{i1}^3,Y_{i0}^1,X_i), \psi_\theta^{1|1}(Y_{i1}^3,Y_{i0}^1,X_i)$. As a result, we further conclude that $\mathcal{T}_*^{\perp}=\text{span}\left(\{\psi_\theta^{0|0}(Y_{i1}^3,Y_{i0}^1,X_i),\psi_\theta^{1|1}(Y_{i1}^3,Y_{i0}^1,X_i)\}\right)$. Hence, $\psi_\theta^{eff}(Y_1^3,X)\in\mathcal{T}_*^{\perp}$ since it is a linear combination of $\psi_\theta^{0|0}(Y_{i1}^3,Y_{i0}^1,X_i)$ and $\psi_\theta^{1|1}(Y_{i1}^3,Y_{i0}^1,X_i)$. Finally since $\mathcal{T}_*^{\perp}\subset\mathcal{T}^{\perp}$, $\psi_\theta^{eff}(Y_1^3,X)\in\mathcal{T}^{\perp}$.

D) Verification of condition ii) $S_\theta-\psi_\theta^{eff}(Y_1^3,x)\in\mathcal{T}$

To check condition ii) $S_\theta-\psi_\theta^{eff}(Y_1^3,x)\in\mathcal{T}$, we will verify the equivalent condition that for any element $g\in\mathcal{T}^{\perp}$, $\mathbb{E}\left[\left(S_\theta-\psi_\theta^{eff}(Y_1^3,x)\right)g(Y_1,Y_2,Y_3,x)|x\right]=0$. Given our characterization of $\mathcal{T}^{\perp}$, it is equivalent to verify that $\forall k\in\{0,1\}$,
$\mathbb{E}\left[\left(S_\theta-\psi_\theta^{eff}(Y_1^3,x)\right)\psi_\theta^{k|k}(Y_1^3,Y_0^1,x)|x\right]=0$

D)1) $S_\gamma-\psi_\gamma^{eff}(Y_1^3,x)\perp\psi_\theta^{0|0}(Y_1^3,Y_0^1,x)$

Let $\Delta_\gamma^{0|0}=(S_\gamma-\psi_\gamma^{eff}(Y_1^3,Y_{i0}^1,x))\psi_\theta^{0|0}(Y_1^3,Y_0^1,x)$. It is tedious but straightforward to show that

$$\begin{aligned}
\Delta_\gamma^{0|0} &= \Delta_{\gamma,1}^{0|0}+\Delta_{\gamma,2}^{0|0}+\Delta_{\gamma,3}^{0|0}+\Delta_{\gamma,4}^{0|0}+\Delta_{\gamma,5}^{0|0}\\
\Delta_{\gamma,1}^{0|0} &= (1-C_{101}(x_2))e^{x_{21}'\beta+\gamma}Y_1(1-Y_2)Y_3 - C_{100}(x_2)e^{x_{31}'\beta}Y_1(1-Y_2)(1-Y_3)\\
\Delta_{\gamma,2}^{0|0} &= -(1-C_{011}(x_3))(1-Y_1)Y_2Y_3 + C_{010}(x_3)(1-Y_1)Y_2(1-Y_3)\\
\Delta_{\gamma,3}^{0|0} &= \Omega_{11}(x)(e^{x_{23}'\beta}-1)^2(1-Y_1)(1-Y_2)Y_3 + \Omega_{11}(x)e^{2x_{21}'\beta+2\gamma}Y_1(1-Y_2)Y_3\\
&\quad + \Omega_{11}(x)e^{2X_{31}'\beta}Y_1(1-Y_2)(1-Y_3) + \Omega_{11}(x)(1-Y_1)Y_2\\
\Delta_{\gamma,4}^{0|0} &= -\Omega_{12}(x)e^{x_{21}'\beta+\gamma}Y_1(1-Y_2)Y_3 - \Omega_{12}(x)e^{x_{31}'\beta}Y_1(1-Y_2)(1-Y_3)\\
&\quad - \Omega_{12}(x)e^{x_{12}'\beta}(1-Y_1)Y_2(1-Y_3) - \Omega_{12}(x)e^{x_{13}'\beta-\gamma}(1-Y_1)Y_2Y_3
\end{aligned}$$

$$\Delta^{0|0}_{\gamma,5} = -e^{X'_{21}\beta+\gamma}Y_1(1-Y_2)Y_3$$

We then note that

$$\begin{aligned}
\mathbb{E}\left[\Delta^{0|0}_{\gamma,1}|x\right] &= \int \frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_3\beta+a}}{1+e^{x'_3\beta+a}}e^{x'_{21}\beta+\gamma}\pi(a|x)da \\
&- \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{x'_3\beta+a}}e^{x'_{31}\beta}\pi(a|x)da \\
&= \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_3\beta+a}}{1+e^{x'_3\beta+a}}\pi(a|x)da \\
&- \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_3\beta+a}}{1+e^{x'_3\beta+a}}\pi(a|x)da \\
&= 0
\end{aligned}$$

and by a similar calculation $\mathbb{E}\left[\Delta^{0|0}_{\gamma,2}|x\right] = 0$. Next, we immediately have

$$\begin{aligned}
\mathbb{E}\left[\Delta^{0|0}_{\gamma,3}|x\right] &= \Omega_{11}(x)\Sigma_{11}(x) \\
\mathbb{E}\left[\Delta^{0|0}_{\gamma,4}|x\right] &= \Omega_{12}(x)\Sigma_{12}(x) \\
\mathbb{E}\left[\Delta^{0|0}_{\gamma,5}|x\right] &= -e^{X'_{21}\beta+\gamma}P_{101}(x)
\end{aligned}$$

and hence,

$$\Delta^{0|0}_{\gamma} = \Omega_{11}(x)\Sigma_{11}(X) + \Omega_{12}(x)\Sigma_{12}(x) - e^{x'_{21}\beta+\gamma}P_{101}(x) = D_{11}(x) - D_{11}(x) = 0$$

D)2) $S_\gamma - \psi^{eff}_\gamma(Y^3_1, x) \perp \psi^{1|1}_\theta(Y^3_1, Y^1_0, x)$

Let $\Delta^{1|1}_\gamma = (S_\gamma - \psi^{eff}_\gamma(Y^3_1, Y^1_{i0}, x))\psi^{1|1}_\theta(Y^3_1, Y^1_0, x)$. It can be decomposed as follows

$$\begin{aligned}
\Delta^{1|1}_\gamma &= \Delta^{1|1}_{\gamma,1} + \Delta^{1|1}_{\gamma,2} + \Delta^{1|1}_{\gamma,3} + \Delta^{1|1}_{\gamma,4} + \Delta^{1|1}_{\gamma,5} \\
\Delta^{1|1}_{\gamma,1} &= -(e^{X'_{32}\beta}-1)C_{1,1,0}(x_2)Y_1Y_2(1-Y_3) \\
&\quad + C_{1,0,1}(x_2)Y_1(1-Y_2)Y_3 + C_{1,0,0}(x_2)Y_1(1-Y_2)(1-Y_3) \\
\Delta^{1|1}_{\gamma,2} &= +(e^{X'_{32}\beta}-1)(1-C_{1,1,0}(x_3))Y_1Y_2(1-Y_3) - e^{x'_{12}\beta}C_{0,1,0}(x_3)(1-Y_1)Y_2(1-Y_3) \\
&\quad - e^{x'_{13}\beta-\gamma}C_{0,1,1}(x_3)(1-Y_1)Y_2Y_3 \\
\Delta^{1|1}_{\gamma,3} &= -\Omega_{11}(x)e^{x'_{21}\beta+\gamma}Y_1(1-Y_2)Y_3 - \Omega_{11}(x)e^{x'_{31}\beta}Y_1(1-Y_2)(1-Y_3) \\
&\quad - \Omega_{11}(x)e^{X'_{12}\beta}(1-Y_1)Y_2(1-Y_3) - \Omega_{11}(x)e^{x'_{13}\beta-\gamma}(1-Y_1)Y_2Y_3 \\
\Delta^{1|1}_{\gamma,4} &= +\Omega_{12}(x)(e^{X'_{32}\beta}-1)^2Y_1Y_2(1-Y_3) + \Omega_{12}(x)e^{2x'_{12}\beta}(1-Y_1)Y_2(1-Y_3) \\
&\quad + \Omega_{12}(x)e^{2x'_{13}\beta-2\gamma}(1-Y_1)Y_2Y_3 + \Omega_{12}(x)Y_1(1-Y_2)
\end{aligned}$$

$\Delta^{1|1}_{\gamma,5} = e^{x'_{13}\beta-\gamma}(1-Y_1)Y_2Y_3$

First, we have

$$\begin{aligned}
\mathbb{E}\left[\Delta^{1|1}_{\gamma,1}|x\right] &= -\int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}}(e^{x'_{32}\beta}-1)\pi(a|x)da \\
&\quad + \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\pi(a|x)da \\
&= -\int \frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{\gamma+x'_3\beta+a}}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&\quad + \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&\quad + \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\pi(a|x)da \\
&= +\int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&\quad + \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&= \mathbb{E}[Y_1Y_2(1-Y_3)|Y_0=0,x]
\end{aligned}$$

By a very similar calculation, $\mathbb{E}\left[\Delta^{1|1}_{\gamma,2}|x\right] = -\mathbb{E}[Y_1Y_2(1-Y_3)|Y_0=0,x]$. Then,

$$\begin{aligned}
\mathbb{E}\left[\Delta^{1|1}_{\gamma,3}|x\right] &= \Omega_{11}(x)\Sigma_{12}(x) \\
\mathbb{E}\left[\Delta^{1|1}_{\gamma,4}|x\right] &= \Omega_{12}(x)\Sigma_{22}(x) \\
\mathbb{E}\left[\Delta^{1|1}_{\gamma,5}|x\right] &= +e^{x'_{13}\beta-\gamma}P_{011}(x)
\end{aligned}$$

It follows that

$$\mathbb{E}\left[\Delta^{1|1}_{\gamma}|x\right] = \Omega_{11}(x)\Sigma_{12}(x) + \Omega_{12}(x)\Sigma_{22}(x) + e^{x'_{13}\beta-\gamma}P_{011}(x) = D_{21}(x) - D_{21}(x) = 0$$

D)3) $S_\beta - \psi^{eff}_\beta(Y^3_1,x) \perp \psi^{0|0}_\theta(Y^3_1,Y^1_0,x)$

Fix $j \in \{2,\ldots,K+1\}$. Let $\Delta^{0|0}_{\beta_{j-1}} = (S_{\beta_{j-1}} - \psi^{eff}_{\beta_{j-1}}(Y^3_1,Y^1_{i0},x))\psi^{0|0}_\theta(Y^3_1,Y^1_0,x)$. Tedious calculations and rearrangements lead to the following decomposition:

$$\Delta^{0|0}_{\beta_{j-1}} = \Delta^{0|0}_{\beta_{j-1},1} + \Delta^{0|0}_{\beta_{j-1},2} + \Delta^{0|0}_{\beta_{j-1}}(x_1) + \Delta^{0|0}_{\beta_{j-1}}(x_2) + \Delta^{0|0}_{\beta_{j-1}}(x_3)$$

where

$$\Delta^{0|0}_{\beta_{j-1}}(x_1) = \Delta^{0|0}_{\beta_{j-1},1}(x_1) + \Delta^{0|0}_{\beta_{j-1},2}(x_1)$$

$$\begin{aligned}
\Delta^{0|0}_{\beta_{j-1},1}(x_1) &= -(e^{x'_{23}\beta}-1)x_{1,j-1}B_{001}(x_1)(1-Y_1)(1-Y_2)Y_3 \\
&\quad - e^{x'_{21}\beta+\gamma}x_{1,j-1}B_{101}(x_1)Y_1(1-Y_2)Y_3 \\
&\quad - e^{x'_{31}\beta}x_{1,j-1}B_{100}(x_1)Y_1(1-Y_2)(1-Y_3) \\
&\quad + x_{1,j-1}B_{011}(x_1)(1-Y_1)Y_2Y_3 + x_{1,j-1}B_{010}(x_1)(1-Y_1)Y_2(1-Y_3) \\
\Delta^{0|0}_{\beta_{j-1},2}(x_1) &= e^{x'_{21}\beta+\gamma}x_{1,j-1}(Y_1(1-Y_2)Y_3 + e^{x'_{31}\beta}x_{1,j-1}Y_1(1-Y_2)(1-Y_3)
\end{aligned}$$

and

$$\begin{aligned}
\Delta^{0|0}_{\beta_{j-1}}(x_2) &= \Delta^{0|0}_{\beta_{j-1},1}(x_2) + \Delta^{0|0}_{\beta_{j-1},2}(x_2) + \Delta^{0|0}_{\beta_{j-1},3}(x_2) \\
\Delta^{0|0}_{\beta_{j-1},1}(x_2) &= e^{x'_{23}\beta}x_{2,j-1}(1-B_{001}(x_2))(1-Y_1)(1-Y_2)Y_3 \\
&\quad + x_{2,j-1}B_{001}(x_2)(1-Y_1)(1-Y_2)Y_3 \\
&\quad - x_{2,j-1}(1-B_{011}(x_2))(1-Y_1)Y_2Y_3 - x_{2,j-1}(1-B_{010}(x_2))(1-Y_1)Y_2(1-Y_3) \\
\Delta^{0|0}_{\beta_{j-1},2}(x_2) &= +e^{x'_{21}\beta+\gamma}x_{2,j-1}(1-C_{101}(x_2))Y_1(1-Y_2)Y_3 \\
&\quad - e^{x'_{31}\beta}x_{2,j-1}C_{100}(x_2)Y_1(1-Y_2)(1-Y_3) \\
\Delta^{0|0}_{\beta_{j-1},3}(x_2) &= -e^{x'_{23}\beta}x_{2,j-1}(1-Y_1)(1-Y_2)Y_3 - e^{x'_{21}\beta+\gamma}x_{2,j-1}Y_1(1-Y_2)Y_3
\end{aligned}$$

and

$$\begin{aligned}
\Delta^{0|0}_{\beta_{j-1}}(x_3) &= \Delta^{0|0}_{\beta_{j-1},1}(x_3) + \Delta^{0|0}_{\beta_{j-1},2}(x_3) + \Delta^{0|0}_{\beta_{j-1},3}(x_3) \\
\Delta^{0|0}_{\beta_{j-1},1}(x_3) &= -(e^{x'_{23}\beta}-1)x_{3,j-1}B_{001}(x_3)(1-Y_1)(1-Y_2)Y_3 - x_{3,j-1}(1-Y_1)(1-Y_2)Y_3 \\
&\quad + e^{x'_{21}\beta+\gamma}x_{3,j-1}(1-B_{101}(x_3))Y_1(1-Y_2)Y_3 \\
&\quad + e^{x'_{31}\beta}x_{3,j-1}(1-B_{100}(x_3))Y_1(1-Y_2)(1-Y_3) \\
\Delta^{0|0}_{\beta_{j-1},2}(x_3) &= -x_{3,j-1}(1-C_{011}(x_3))(1-Y_1)Y_2Y_3 + x_{3,j-1}C_{010}(x_3)(1-Y_1)Y_2(1-Y_3) \\
\Delta^{0|0}_{\beta_{j-1},3}(x_3) &= e^{x'_{23}\beta}x_{3,j-1}(1-Y_1)(1-Y_2)Y_3 - e^{x'_{31}\beta}x_{3,j-1}Y_1(1-Y_2)(1-Y_3)
\end{aligned}$$

and last

$$\begin{aligned}
\Delta^{0|0}_{\beta_{j-1},1} &= +\Omega_{j1}(x)(e^{x'_{23}\beta}-1)^2(1-Y_1)(1-Y_2)Y_3 + \Omega_{j1}(x)e^{2x'_{21}\beta+2\gamma}Y_1(1-Y_2)Y_3 \\
&\quad + \Omega_{j1}(x)e^{2x'_{31}\beta}Y_1(1-Y_2)(1-Y_3) + \Omega_{j1}(x)(1-Y_1)Y_2 \\
\Delta^{0|0}_{\beta_{j-1},2} &= -\Omega_{j2}(x)e^{x'_{21}\beta+\gamma}Y_1(1-Y_2)Y_3 - \Omega_{j2}(x)e^{x'_{31}\beta}Y_1(1-Y_2)(1-Y_3) \\
&\quad - \Omega_{j2}(x)e^{x'_{12}\beta}(1-Y_1)Y_2(1-Y_3) - \Omega_{j2}(x)e^{x'_{13}\beta-\gamma}(1-Y_1)Y_2Y_3
\end{aligned}$$

Starting first with the terms in "x_1", we have:

$$\frac{1}{x_{1,j-1}}\mathbb{E}[\Delta^{0|0}_{\beta_{j-1},1}(x_1)|x] = \mathbb{E}\left[\frac{e^{x'_1\beta+A}}{1+e^{x'_1\beta+A}}\mathbb{E}\left[-\psi^{0|0}_\theta(Y_1^3,Y_0^1,x)|x,A\right]|x\right] = 0$$

$$\mathbb{E}[\Delta^{0|0}_{\beta_{j-1},2}(x_1)|x] = e^{x'_{21}\beta+\gamma}x_{1,j-1}P_{101}(x) + e^{x'_{31}\beta}x_{1,j-1}P_{100}(x)$$

Next, for the terms in "x_2", we have:

$$\begin{aligned}
\frac{1}{x_{2,j-1}}\mathbb{E}\left[\Delta_{\beta,1}(x_2)|x\right] &= \int \frac{1}{1+e^{x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{x'_2\beta+a}}\frac{e^{x'_3\beta+a}}{1+e^{x'_3\beta+a}}e^{x'_{23}\beta}\pi(a|x)da \\
&+ \int \frac{e^{x'_2\beta+a}}{1+e^{x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{x'_2\beta+a}}\frac{e^{x'_3\beta+a}}{1+e^{x'_3\beta+a}}\pi(a|x)da \\
&- \int \frac{1}{1+e^{x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{e^{x'_2\beta+a}}{1+e^{x'_2\beta+a}}\pi(a|x)da \\
&= \int \frac{1}{1+e^{x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{e^{x'_2\beta+a}}{1+e^{x'_2\beta+a}}\pi(a|x)da \\
&- \int \frac{1}{1+e^{x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{e^{x'_2\beta+a}}{1+e^{x'_2\beta+a}}\pi(a|x)da \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\frac{1}{x_{2,j-1}}E\left[\Delta_{\beta,2}(x_2)|x\right] &= \int \frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_3\beta+a}}{1+e^{x'_3\beta+a}}e^{x'_{21}\beta+\gamma}\pi(a|x)da \\
&- \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{x'_3\beta+a}}e^{x'_{31}\beta}\pi(a|x)da \\
&= \int \frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_3\beta+a}}{1+e^{x'_3\beta+a}}\pi(a|x)da \\
&- \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_3\beta+a}}{1+e^{x'_3\beta+a}}\pi(a|x)da \\
&= 0
\end{aligned}$$

$$\mathbb{E}\left[\Delta^{0|0}_{\beta_{j-1},3}(x_2)|x\right] = -e^{x'_{23}\beta}x_{2,j-1}P_{001}(x) - e^{x'_{21}\beta+\gamma}x_{2,j-1}P_{101}(x)$$

By the same token, for the terms in "x_3", one arrives at
$\mathbb{E}\left[\Delta^{0|0}_{\beta_{j-1},1}(x_3)|x\right] = \mathbb{E}\left[\Delta^{0|0}_{\beta_{j-1},2}(x_3)|x\right] = 0$ and

$$\begin{aligned}
\mathbb{E}\left[\Delta^{0|0}_{\beta_{j-1},1}(x_3)|x\right] &= \mathbb{E}\left[\Delta^{0|0}_{\beta_{j-1},2}(x_3)|x\right] = 0 \\
\mathbb{E}\left[\Delta^{0|0}_{\beta_{j-1},3}(x_3)|x\right] &= e^{x'_{23}\beta}x_{3,j-1}P_{001}(x) - e^{x'_{31}\beta}x_{3,j-1}P_{100}(x)
\end{aligned}$$

Finally, $\mathbb{E}[\Delta^{0|0}_{\beta_{j-1},1}|x] = \Omega_{j,1}(x)\Sigma_{11}(x), \mathbb{E}[\Delta^{0|0}_{\beta_{j-1},2}|x] = \Omega_{j,2}(x)\Sigma_{12}(x)$. Collecting terms, we get

$$\mathbb{E}\left[\Delta^{0|0}_{\beta_{j-1}}|x\right] = e^{x'_{21}\beta+\gamma}x_{1,j-1}P_{101}(x) + e^{x'_{31}\beta}x_{1,j-1}P_{100}(x)$$

$$
\begin{aligned}
&- e^{x'_{23}\beta} x_{2,j-1} P_{001}(x) - e^{x'_{21}\beta+\gamma} x_{2,j-1} P_{101}(x) \\
&e^{x'_{23}\beta} x_{3,j-1} P_{001}(x) - e^{x'_{31}\beta} x_{3,j-1} P_{100}(x) + \Omega_{j1}(x)\Sigma_{11}(x) + \Omega_{j2}(x)\Sigma_{12}(x) \\
&= -D_{1j}(x) + D_{1j}(x) \\
&= 0
\end{aligned}
$$

This is of course valid for all slope parameters β_j and hence $S_\beta - \psi_\beta^{eff}(Y_1^3, x) \perp \psi_\theta^{0|0}(Y_1^3, Y_0^1, x)$

D)4) $S_\beta - \psi_\beta^{eff}(Y_1^3, x) \perp \psi_\theta^{1|1}(Y_1^3, Y_0^1, x)$

Fix $j \in \{2, \ldots, K+1\}$. Let $\Delta^{1|1}_{\beta_{j-1}} = (S_{\beta_{j-1}} - \psi^{eff}_{\beta_{j-1}}(Y_1^3, Y_{i0}^1, x))\psi_\theta^{1|1}(Y_1^3, Y_0^1, x)$. A last set of lengthy calculations and rearrangements lead to the following decomposition:

$$
\Delta^{1|1}_{\beta_{j-1}} = \Delta^{1|1}_{\beta_{j-1},1} + \Delta^{1|1}_{\beta_{j-1},2} + \Delta^{1|1}_{\beta_{j-1}}(x_1) + \Delta^{1|1}_{\beta_{j-1}}(x_2) + \Delta^{1|1}_{\beta_{j-1}}(x_3)
$$

where

$$
\begin{aligned}
\Delta^{1|1}_{\beta_{j-1}}(x_1) &= \Delta^{1|1}_{\beta_{j-1},1}(x_1) + \Delta^{1|1}_{\beta_{j-1},2}(x_1) \\
\Delta^{1|1}_{\beta_{j-1},1}(x_1) &= +(e^{x'_{32}\beta} - 1) x_{1,j-1}(1 - B_{110}(x_1)) Y_1 Y_2 (1 - Y_3) \\
&\quad + e^{x'_{12}\beta} x_{1,j-1}(1 - B_{010}(x_1))(1 - Y_1) Y_2 (1 - Y_3) \\
&\quad + e^{x'_{13}\beta-\gamma} x_{1,j-1}(1 - B_{011}(x_1))(1 - Y_1) Y_2 Y_3 \\
&\quad - x_{1,j-1}(1 - B_{101}(x_1)) Y_1 (1 - Y_2) Y_3 - x_{1,j-1}(1 - B_{100}(x_1)) Y_1 (1 - Y_2)(1 - Y_3) \\
\Delta^{1|1}_{\beta_{j-1},2}(x_1) &= -e^{x'_{12}\beta} x_{1,j-1}(1 - Y_1) Y_2 (1 - Y_3) - e^{x'_{13}\beta-\gamma} x_{1,j-1}(1 - Y_1) Y_2 Y_3
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta^{1|1}_{\beta_{j-1}}(x_2) &= \Delta^{1|1}_{\beta_{j-1},1}(x_2) + \Delta^{1|1}_{\beta_{j-1},2}(x_2) + \Delta^{1|1}_{\beta_{j-1},3}(x_2) \\
\Delta^{1|1}_{\beta_{j-1},1}(x_2) &= -e^{x'_{32}\beta} x_{2,j-1} C_{110}(x_2) Y_1 Y_2 (1 - Y_3) - x_{2,j-1}(1 - C_{110}(x_2)) Y_1 Y_2 (1 - Y_3) \\
&\quad + x_{2,j-1} C_{101}(x_2) Y_1 (1 - Y_2) Y_3 + x_{2,j-1} C_{100}(x_2) Y_1 (1 - Y_2)(1 - Y_3) \\
\Delta^{1|1}_{\beta_{j-1},2}(x_2) &= -e^{x'_{12}\beta} x_{2,j-1} B_{010}(x_2)(1 - Y_1) Y_2 (1 - Y_3) \\
&\quad + e^{x'_{13}\beta-\gamma} x_{2,j-1}(1 - B_{011}(x_2))(1 - Y_1) Y_2 Y_3 \\
\Delta^{1|1}_{\beta_{j-1},3}(x_2) &= e^{x'_{32}\beta} x_{2,j-1} Y_1 Y_2 (1 - Y_3) + e^{x'_{12}\beta} x_{2,j-1}(1 - Y_1) Y_2 (1 - Y_3)
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta^{1|1}_{\beta_{j-1}}(x_3) &= \Delta^{1|1}_{\beta_{j-1},1}(x_3) + \Delta^{1|1}_{\beta_{j-1},2}(x_3) + \Delta^{1|1}_{\beta_{j-1},3}(x_3) \\
\Delta^{1|1}_{\beta_{j-1},1}(x_3) &= +e^{x'_{32}\beta} x_{3,j-1}(1 - C_{110}(x_3)) Y_1 Y_2 (1 - Y_3) + x_{3,j-1} C_{110}(x_3) Y_1 Y_2 (1 - Y_3) \\
&\quad - e^{x'_{12}\beta} x_{3,j-1} C_{010}(x_3)(1 - Y_1) Y_2 (1 - Y_3) - e^{x'_{13}\beta-\gamma} x_{3,j-1} C_{011}(x_3)(1 - Y_1) Y_2 Y_3 \\
\Delta^{1|1}_{\beta_{j-1},2}(x_3) &= -x_{3,j-1}(1 - B_{101}(x_3)) Y_1 (1 - Y_2) Y_3 + x_{3,j-1} B_{100}(x_3) Y_1 (1 - Y_2)(1 - Y_3)
\end{aligned}
$$

$$\Delta^{1|1}_{\beta_{j-1},3}(x_3) = e^{x'_{13}\beta-\gamma}x_{3,j-1}(1-Y_1)Y_2Y_3 - e^{x'_{32}\beta}x_{3,j-1}Y_1Y_2(1-Y_3)$$

and last

$$\begin{aligned}
\Delta^{1|1}_{\beta_{j-1},1} &= -\Omega_{j1}(x)e^{x'_{21}\beta+\gamma}Y_1(1-Y_2)Y_3 - \Omega_{j1}(x)e^{x'_{31}\beta}Y_1(1-Y_2)(1-Y_3) \\
&\quad - \Omega_{j1}(x)e^{x'_{12}\beta}(1-Y_1)Y_2(1-Y_3) - \Omega_{j1}(x)e^{x'_{13}\beta-\gamma}(1-Y_1)Y_2Y_3 \\
\Delta^{1|1}_{\beta_{j-1},2} &= +\Omega_{j2}(x)(e^{x'_{32}\beta}-1)^2Y_1Y_2(1-Y_3) + \Omega_{j2}(x)e^{2x'_{12}\beta}(1-Y_1)Y_2(1-Y_3) \\
&\quad + \Omega_{j2}(x)e^{2x'_{13}\beta-2\gamma}(1-Y_1)Y_2Y_3 + \Omega_{j2}(x)Y_1(1-Y_2)
\end{aligned}$$

Starting first with the terms in "x_1", we have:

$$\begin{aligned}
\frac{1}{x_{1,j-1}}\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},1}(x_1)|x\right] &= \mathbb{E}\left[\frac{1}{1+e^{x'_1\beta+A}}\mathbb{E}\left[\psi^{1|1}_\theta(Y^3_{i1},Y^1_{i0},X_i)|x,A\right]|x\right] = 0 \\
\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},2}(x_1)|x\right] &= -e^{x'_{12}\beta}x_{1,j-1}P_{010}(x) - e^{x'_{13}\beta-\gamma}x_{1,j-1}P_{011}(x)
\end{aligned}$$

For the terms in "x_2"

$$\begin{aligned}
\frac{1}{x_{2,j-1}}\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},1}(x_2)|x\right] &= -\int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}} \\
&\quad \times e^{x'_{32}\beta}\pi(a|x)da \\
&\quad - \int \frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&\quad + \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\pi(a|x)da \\
&= -\int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{\gamma+x'_3\beta+a}}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&\quad + \int \frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{1}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{\gamma+x'_3\beta+a}}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&= 0 \\
\frac{1}{x_{2,j-1}}\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},2}(x_2)|x\right] &= -\int \frac{e^{x'_2\beta+a}}{1+e^{x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}}e^{x'_{12}\beta}\pi(a|x)da \\
&\quad + \int \frac{1}{1+e^{x'_2\beta+a}}\frac{1}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{e^{\gamma+x'_3\beta+a}}{1+e^{\gamma+x'_3\beta+a}}e^{x'_{13}\beta-\gamma}\pi(a|x)da \\
&= -\int \frac{1}{1+e^{x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&\quad + \int \frac{1}{1+e^{x'_2\beta+a}}\frac{e^{x'_1\beta+a}}{1+e^{x'_1\beta+a}}\frac{e^{\gamma+x'_2\beta+a}}{1+e^{\gamma+x'_2\beta+a}}\frac{1}{1+e^{\gamma+x'_3\beta+a}}\pi(a|x)da \\
&= 0
\end{aligned}$$

$$\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},3}(x_2)|x\right] = e^{x'_{32}\beta}x_{2,j-1}P_{110}(x) + e^{x'_{12}\beta}x_{2,j-1}P_{010}(x)$$

Similar calculations for the terms in "x_3" yield $\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},1}(x_3)|x\right] = \mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},2}(x_3)|x\right] = 0$ and $\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},3}(x_3)|x\right] = e^{x'_{13}\beta-\gamma}x_{3,j-1}P_{011}(x) - e^{x'_{32}\beta}x_{3,j-1}P_{110}(x)$. Finally,

$$\begin{aligned}
\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},1}|x\right] &= -\Omega_{j1}(x)e^{x'_{21}\beta+\gamma}P_{101}(x) - \Omega_{j1}(x)e^{x'_{31}\beta}P_{100}(x) \\
&\quad - \Omega_{j1}(x)e^{x'_{12}\beta}P_{010}(x) - \Omega_{j1}(x)e^{x'_{13}\beta-\gamma}P_{011}(x) \\
&= \Omega_{j1}(x)\Sigma_{12}(x) \\
\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1},2}|x\right] &= +\Omega_{j2}(x)(e^{x'_{32}\beta}-1)^2P_{110}(x) + \Omega_{j2}(x)e^{2x'_{12}\beta}P_{010}(x) \\
&\quad + \Omega_{j2}(x)e^{2x'_{13}\beta-2\gamma}P_{011}(x) + \Omega_{j2}(x)P_{10}(x) \\
&= \Omega_{j2}(x)\Sigma_{22}(x)
\end{aligned}$$

Putting the different pieces together, we ultimately obtain

$$\begin{aligned}
\mathbb{E}\left[\Delta^{1|1}_{\beta_{j-1}}|x\right] &= -e^{x'_{12}\beta}x_{1,j-1}P_{010}(x) - e^{x'_{13}\beta-\gamma}x_{1,j-1}P_{011}(x) \\
&\quad + e^{x'_{32}\beta}x_{2,j-1}P_{110}(x) + e^{x'_{12}\beta}x_{2,j-1}P_{010}(x) \\
&\quad + e^{x'_{13}\beta-\gamma}x_{3,j-1}P_{011}(x) - e^{x'_{32}\beta}x_{3,j-1}P_{110}(x) \\
&\quad + \Omega_{j1}(x)\Sigma_{12}(x) + \Omega_{j2}(x)\Sigma_{22}(x) \\
&= -D_{2j}(x) + D_{2j}(x) \\
&= 0
\end{aligned}$$

This is of course valid for all slope parameters β_j and hence $S_\beta - \psi^{eff}_\beta(Y_1^3, x) \perp \psi^{1|1}_\theta(Y_1^3, Y_0^1, x)$

E) Conclusion

Having verified all the conditions of Theorem 3.2 in Newey (1990) for the initial condition $Y_0 = 0$, we conclude that in that case $\psi^{eff}_\theta(Y_1^3, X)$ is the efficient score of the AR(1) model. The semiparametric efficiency bound is given by $\mathbb{E}\left[D(X)'\Sigma(X)^{-1}D(X)\right]^{-1}$. Symmetric results can be shown to hold for the case $Y_0 = 1$.

Chapter 2

Identification in a Binary Choice Panel Data Model with a Predetermined Covariate[1]

2.1 Introduction

Empirical researchers utilizing panel data generally maintain the assumption that covariates are strictly exogenous: realized values of past, current, and future explanatory variables are independent of the time-varying structural disturbances or "shocks".[2] In many settings this assumption is unrealistic. If the covariate is a policy, choice or dynamic state variable, then agents may adjust its level in response to past shocks (as when, for example, a firm adjusts its current capital expenditures in response to past productivity shocks).

When strict exogeneity is untenable, *sequential exogeneity* – sometimes called *predeterminedness* – may be palatable. A predetermined covariate varies independently of current and future time-varying shocks, but general *feedback*, or dependence on past shocks, is allowed. Assumptions of this type play an important role in, for example, production function estimation (Olley and Pakes, 1996, Blundell and Bond, 2000).

In two seminal papers, Arellano and Bond (1991) and Arellano and Bover (1995), Manuel Arellano and his collaborators presented foundational analyses of questions of identification, estimation, efficiency and specification testing in linear panel data models with feedback. Today such models are both well-understood and widely-used (see Arellano (2003) for a textbook review).

In contrast, the properties of nonlinear models with feedback are much less well-understood. In this chapter we study binary choice. Most existing work in this area focuses on the case where the covariate is either strictly exogenous or a lagged outcome. Under strict exo-

[1] This chapter is joint work with Stéphane Bonhomme and Bryan Graham.

[2] Dependence between the covariates and the time-invariant heterogeneity – the so-called "fixed effects" – is, of course, allowed.

geneity, Rasch (1960) and Andersen (1970) show that the coefficient on the covariate is point-identified using two periods of data when shocks are logistic. Chamberlain (2010) provides conditions under which the logit case is the only one admitting point-identification with two periods (Davezies et al. (2020) provide extensions of this result to the case of $T > 2$). In the dynamic case, where the covariate is a lagged outcome, Cox (1958b), Chamberlain (1985a) and Honoré and Kyriazidou (2000) derive conditions for point-identification of the coefficient on the lagged outcome in the logit case, while Honoré and Tamer (2006) show how to compute bounds on coefficients for probit and other models.

Results for binary choice panel models with predetermined covariates are limited. Chamberlain (2022) studies identification and semiparametric efficiency bounds in a class of nonlinear panel data models with feedback; he provides both positive and negative results. In an hitherto unpublished section of an early draft of that paper (Chamberlain, 1993), he proves that the coefficient on a lagged outcome is not point-identified in a dynamic logit model when only three periods of outcome data are available. Arellano and Carrasco (2003a) and Honoré and Lewbel (2002) study binary choice models with predetermined covariates. Arellano and Carrasco (2003a) assume that the dependence between the time-invariant heterogeneity and the covariates is fully characterized by its conditional mean given current and lagged covariates. Honoré and Lewbel (2002) assume that one of the covariates is independent of the individual effects conditional on the other covariates. In a recent contribution, Pigini and Bartolucci (2022) show that one can accommodate specific forms of feedback while maintaining point-identification in binary choice models with pretermined covariates.[3]

In what follows we pose two questions. First, under what conditions is the coefficient on a predetermined covariate in a binary choice panel data model point-identified? Second, when the coefficient is only set-identified, how extreme is the failure of point-identification; i.e., what is the width of the identified set?

Our analyses leave the dependence between the (time-invariant) unit-specific heterogeneity and the covariates unrestricted. We focus on the special case of a single binary predetermined covariate, leaving the feedback process from lagged outcomes, covariates and the unit-specific heterogeneity onto future covariate realizations fully unrestricted. This is a substantial relaxation of the strict exogeneity assumption.

Regarding point-identification, we provide a simple condition on the model which guarantees that point-identification fails when T periods of data are available (and T is fixed). The condition is satisfied in most familiar models of binary choice, including the logit one. This finding contrasts with the prior work on logit models cited above, where point-identification typically holds for a sufficiently long panel. As a notable exception, the exponential binary choice model introduced by Al-Sadoon et al. (2017) does not satisfy our condition. In fact, point-identification holds in that case.

Regarding identified sets, we first show that sharp bounds on the coefficient can be

[3]In this chapter we focus on panel data with a fixed number T of time periods. The large-T literature has also considered models with dynamics and feedback, see for example Carro (2007), Hahn and Kuersteiner (2002), and Fernández-Val (2009).

computed using linear programming techniques. Our method builds on Honoré and Tamer (2006), however, in contrast to their work, we allow for heterogeneous feedback. While the regressor coefficient is our main target parameter, we also derive the identified set for an average partial effect. This set can be computed using linear programming techniques as well.

Second, we numerically compute examples of identified sets. We find that, relative to the strictly exogenous case, allowing for a predetermined covariate tends to increase the width of the identified set. However, our calculations also suggest that the identified set can remain informative under predeterminedness, even in panels with as few as two periods, for both the coefficient and the average partial effect. Finally, as is true under strict exogeneity, the widths of the identified sets decrease quickly as the number of periods increases. These observations are based upon sets computed under a particular data generating processe (DGP). It is possible that identified sets may be larger under certain types of feedback.

The outline of the chapter is as follows. In Section 2.2 we present the model. In Section 2.3 we provide a condition that implies that the common parameter in this model is not point-identified when $T = 2$. In Section 2.4 we show that our condition implies failure of point-identification for all (finite) T. In Section 2.5 we show how to compute identified sets on coefficients and average partial effects, and we report the results of a small set of numerical illustrations. In Section 2.6 we describe potential restrictions one could impose on the feedback process. These restrictions may restore point-identification or shrink the identified set. We conclude in Section 2.7. Proofs are contained in the appendix. Lastly, replication codes are available as supplementary material.

2.2 The model

Available to the econometrician is a random sample of n units, each of which is followed for $T \geq 2$ time periods. We focus on short panels, and keep T fixed. The sampling process asymptotically reveals the joint distribution of $(X_1, \ldots, X_T, Y_1, \ldots, Y_T)$.

For any sequence of random variables Z_t and any non-stochastic sequence z_t, we use the shorthand notation $Z^{t:t+s} = (Z_t', ..., Z_{t+s}')'$ and $z^{t:t+s} = (z_t', ..., z_{t+s}')'$. In addition, we simply denote $Z^t = Z^{1:t}$ and $z^t = z^{1:t}$ when the subsequence starts in the first period.

Let $Y_{it} \in \{0, 1\}$ and $X_{it} \in \{0, 1\}$ denote a binary outcome and a binary covariate, respectively. We assume that

$$\Pr\big(Y_{it} = 1 \,|\, Y_i^{t-1}, X_i^t, \alpha_i; \theta\big) = F(\theta X_{it} + \alpha_i), \quad t = 1, \ldots, T,$$

where $\alpha_i \in \mathcal{S} \subset \mathbb{R}$ is a scalar individual effect, $F(\cdot)$ is a known differentiable cumulative distribution function, and $\theta \in \Theta$ is a scalar parameter.

Let $\pi_{x_1}(\alpha)$ denote the *distribution of heterogeneity* given the initial condition $X_1 = x_1$; i.e., the distribution of $\alpha_i \,|\, X_{i1}$. We leave this distribution unrestricted on $\mathcal{S}$. When $\mathcal{S}$ is a discrete subset of the real line, $\pi_{x_1}(\alpha)$ belongs to the unit simplex on $\mathcal{S}$, however it is

otherwise unrestricted. We denote as Π the collection of all $\pi_{x_1}(\alpha)$, for all $x_1 \in \{0,1\}$ and $\alpha \in \mathcal{S}$.

For each $t \geq 2$, let

$$\Pr\left(X_{it} = 1 \,|\, Y_i^{t-1} = y^{t-1}, X_i^{t-1} = x^{t-1}, \alpha_i = \alpha\right) = G^t_{y^{t-1},x^{t-1}}(\alpha), \quad t = 2, \ldots, T,$$

denote the *feedback process* through which lagged outcomes, past covariates and heterogeneity affect the current covariate. We leave this distribution unrestricted as well. We denote as $G \in \mathcal{G}_T$ the collection of all $G^t_{y^{t-1},x^{t-1}}(\alpha)$, for all $t \in \{2, ..., T\}$, $y^{t-1} \in \{0,1\}^{t-1}$, $x^{t-1} \in \{0,1\}^{t-1}$, and $\alpha \in \mathcal{S}$.

The (integrated) likelihood function conditional on the first period's covariate is

$$\begin{aligned}\Pr\left(Y_i^T = y^T, X_i^{2:T} = x^{2:T} \,|\, X_{i1} = x_1\right) = \int_{\mathcal{S}} & \underbrace{\prod_{t=1}^{T} F(\theta x_t + \alpha)^{y_t}[1 - F(\theta x_t + \alpha)]^{1-y_t}}_{\text{outcomes}} \\ & \times \underbrace{\prod_{t=2}^{T} G^t_{y^{t-1},x^{t-1}}(\alpha)^{x_t}[1 - G^t_{y^{t-1},x^{t-1}}(\alpha)]^{1-x_t}}_{\text{feedback}} \\ & \times \underbrace{\pi_{x_1}(\alpha)}_{\text{heterogeneity}} \, d\mu(\alpha), \end{aligned} \tag{2.1}$$

for some (discrete or continuous) measure μ on $\mathcal{S}$.

A key feature of a model with predetermined covariates is the dependence of the feedback process on lagged outcomes, as reflected in the dependence of G^t on y^{t-1} in (2.1). When this dependence is ruled out, the covariate is strictly exogenous, and the likelihood function simplifies.[4] Dynamic responses of covariates to lagged outcome realizations are central to many economic models, including those where X_{it} is a choice variable, policy, or a dynamic state variable.

For any $(\theta, \pi, G) \in \Theta \times \Pi \times \mathcal{G}_T$, and any $(y^T, x^{2:T}) \in \{0,1\}^{2T-1}$, let $Q_{x_1}(y^T, x^{2:T}; \theta, \pi, G)$ denote the right-hand side of (2.1). Moreover, let $Q_{x_1}(\theta, \pi, G)$ denote the $2^{2T-1} \times 1$ vector collecting all those elements, for all $(y^T, x^{2:T}) \in \{0,1\}^{2T-1}$. Finally, let $Q(\theta, \pi, G)$ denote the $2^{2T} \times 1$ vector stacking $Q_1(\theta, \pi, G)$ and $Q_0(\theta, \pi, G)$. For a given (population) $(\theta, \pi, G) \in \Theta \times \Pi \times \mathcal{G}_T$, we define the *identified set* of θ as

$$\Theta^I = \left\{\widetilde{\theta} \in \Theta \,:\, \exists(\widetilde{\pi}, \widetilde{G}) \in \Pi \times \mathcal{G}_T \,:\, Q(\widetilde{\theta}, \widetilde{\pi}, \widetilde{G}) = Q(\theta, \pi, G)\right\}. \tag{2.2}$$

[4] Under strict exogeneity, the likelihood function factors as

$$\begin{aligned}\Pr\left(Y_i^T = y^T, X_i^{2:T} = x^{2:T} \,|\, X_{i1} = x_1\right) = & \left[\int_{\mathcal{S}} \prod_{t=1}^{T} F(\theta x_t + \alpha)^{y_t}[1 - F(\theta x_t + \alpha)]^{1-y_t} \pi_{x^T}(\alpha) d\mu(\alpha)\right] \\ & \times \Pr\left(X_i^{2:T} = x^{2:T} \,|\, X_{i1} = x_1\right),\end{aligned}$$

where $\pi_{x^T}(\alpha)$ denotes the distribution of heterogeneity given all periods' covariates $x_1, ..., x_T$.

The set in (2.2) includes all $\widetilde{\theta} \in \Theta$ where, for that $\widetilde{\theta}$, it is possible to find a heterogeneity distribution $\widetilde{\pi} \in \Pi$, *and* a feedback process $\widetilde{G} \in \mathcal{G}_T$, such that the resulting conditional likelihood assigns the same probability to each of the 2^{2T-1} possible data outcomes as the true one (given both $X_{i1} = 0$ and $X_{i1} = 1$).

In the first part of the chapter, we provide conditions on the model under which Θ^I is not a singleton. This corresponds to cases where θ is not point-identified. In the second part of the chapter, we report numerical calculations of Θ^I under particular DGPs.

Our focus on θ is motivated by the extensive literature on the identification of coefficients in binary choice models. However, in applications, average effects may also be of interest. In the second part of the chapter, we will also report numerical calculations of identified sets for an average partial effect associated with a change in the binary predetermined covariate.

2.3 Failure of point-identification in two-period panels

We first present an analysis of point-identification in the two-period case, since this leads to simple and transparent calculations. In the next section, we will then generalize this result to accommodate $T \geq 2$ periods.

2.3.1 Assumptions and result

To keep the formal analysis simple, in this section and the next we assume that α_i takes a finite number of values, with known support points.

Assumption 4. *$\mathcal{S} = \{\underline{\alpha}_1, ..., \underline{\alpha}_K\}$, where $\underline{\alpha}_1, ..., \underline{\alpha}_K$ are known, and $\mu = \sum_{k=1}^{K} \delta_{\underline{\alpha}_k}$, where δ_α denotes the Dirac measure at α.*

Assumption 4 makes the model fully parametric. However this is not a limitation as our aim in this section and the next is to derive conditions under which point-identification *fails*. The conditions we provide will require sufficiently many support points.[5]

We rely on the parameterization given by the $2(K-1) \times 1$ vector $\pi = (\pi_1', \pi_0')'$, where, for all $x_1 \in \{0, 1\}$, $\pi_{x_1} = (\pi_{x_1}(\underline{\alpha}_1), \ldots, \pi_{x_1}(\underline{\alpha}_{K-1}))'$ and $\pi_{x_1}(\underline{\alpha}_K) = 1 - \sum_{k=1}^{K-1} \pi_{x_1}(\underline{\alpha}_k)$. The vector $\pi \in \Pi$ is unrestricted, except for the fact that $\pi_{x_1}(\alpha)$, for $\alpha \in \mathcal{S}$, belongs to the unit simplex. This parameterization handles the fact that probability mass functions sum to one.

We next impose the following assumption on the population parameters.

Assumption 5. *$\theta \in \Theta$, $\pi \in \Pi$, and $G \in \mathcal{G}_T$ are all interior, and $F(\theta x + \alpha) \in (0, 1)$ for all $x \in \{0, 1\}$ and $\alpha \in \mathcal{S}$.*

[5] The analysis is essentially unchanged if one instead assumes that $\mu = \sum_{k=1}^{K} \lambda_k \delta_{\underline{\alpha}_k}$, for some $\lambda_k > 0$.

Assumption 5 places restrictions on the underlying parametric binary choice model and heterogeneity distribution. It rules out heterogeneity distributions that induce a point mass of "stayers" (i.e., units with such extreme values of α that they either always take the binary action or they never do).[6] Assumption 5 also rules out the "staggered adoption" design common in difference-in-differences analyses. Exploring the implications of non-interior feedback processes is left for future work.

Finally, we assume that the parameter point is regular in the sense of Rothenberg (1971).

Assumption 6. *(θ, π, G) is a regular point of the Jacobian matrix $\nabla Q(\theta, \pi, G)$, in the sense that the rank of $\nabla Q(\widetilde{\theta}, \widetilde{\pi}, \widetilde{G})$ is constant for all $(\widetilde{\theta}, \widetilde{\pi}, \widetilde{G})$ in an open neighborhood of (θ, π, G).*

The assumption of regularity is standard in the literature on the identification of parametric models (Rothenberg, 1971). If $F(\cdot)$ is analytic, the irregular points of $\nabla Q(\theta, \pi, G)$ (i.e., the points (θ, π, G) such that Assumption 6 is not satisfied) form a set of measure zero (Bekker and Wansbeek, 2001). Thus, Assumption 6 is satisfied almost everywhere in the parameter space in many binary choice models, including the probit and logit ones.

We aim to provide a simple condition under which point-identification of θ fails when $T = 2$. We start by observing that, when $T = 2$, the $2^{2T-1} = 8$ model outcome probabilities given $X_{i1} = x_1$ are

$$Q_{x_1}(\theta, \pi, G) = \begin{pmatrix} \Pr\big(Y_{i2} = 1, X_{i2} = 1, Y_{i1} = 1 \,|\, X_{i1} = x_1; \theta, \pi, G\big) \\ \Pr\big(Y_{i2} = 1, X_{i2} = 1, Y_{i1} = 0 \,|\, X_{i1} = x_1; \theta, \pi, G\big) \\ \Pr\big(Y_{i2} = 1, X_{i2} = 0, Y_{i1} = 1 \,|\, X_{i1} = x_1; \theta, \pi, G\big) \\ \Pr\big(Y_{i2} = 1, X_{i2} = 0, Y_{i1} = 0 \,|\, X_{i1} = x_1; \theta, \pi, G\big) \\ \Pr\big(Y_{i2} = 0, X_{i2} = 1, Y_{i1} = 1 \,|\, X_{i1} = x_1; \theta, \pi, G\big) \\ \Pr\big(Y_{i2} = 0, X_{i2} = 1, Y_{i1} = 0 \,|\, X_{i1} = x_1; \theta, \pi, G\big) \\ \Pr\big(Y_{i2} = 0, X_{i2} = 0, Y_{i1} = 1 \,|\, X_{i1} = x_1; \theta, \pi, G\big) \\ \Pr\big(Y_{i2} = 0, X_{i2} = 0, Y_{i1} = 0 \,|\, X_{i1} = x_1; \theta, \pi, G\big) \end{pmatrix},$$

which, given the structure of the model, coincide with

$$Q_{x_1}(\theta, \pi, G) = \begin{pmatrix} \int_{\mathcal{S}} F(\theta + \alpha) G_{1,x_1}^2(\alpha) F(\theta x_1 + \alpha) \pi_{x_1}(\alpha) d\mu(\alpha) \\ \int_{\mathcal{S}} F(\theta + \alpha) G_{0,x_1}^2(\alpha) [1 - F(\theta x_1 + \alpha)] \pi_{x_1}(\alpha) d\mu(\alpha) \\ \int_{\mathcal{S}} F(\alpha) [1 - G_{1,x_1}^2(\alpha)] F(\theta x_1 + \alpha) \pi_{x_1}(\alpha) d\mu(\alpha) \\ \int_{\mathcal{S}} F(\alpha) [1 - G_{0,x_1}^2(\alpha)] [1 - F(\theta x_1 + \alpha)] \pi_{x_1}(\alpha) d\mu(\alpha) \\ \int_{\mathcal{S}} \big[1 - F(\theta + \alpha)\big] G_{1,x_1}^2(\alpha) F(\theta x_1 + \alpha) \pi_{x_1}(\alpha) d\mu(\alpha) \\ \int_{\mathcal{S}} \big[1 - F(\theta + \alpha)\big] G_{0,x_1}^2(\alpha) [1 - F(\theta x_1 + \alpha)] \pi_{x_1}(\alpha) d\mu(\alpha) \\ \int_{\mathcal{S}} \big[1 - F(\alpha)\big] [1 - G_{1,x_1}^2(\alpha)] F(\theta x_1 + \alpha) \pi_{x_1}(\alpha) d\mu(\alpha) \\ \int_{\mathcal{S}} \big[1 - F(\alpha)\big] [1 - G_{0,x_1}^2(\alpha)] [1 - F(\theta x_1 + \alpha)] \pi_{x_1}(\alpha) d\mu(\alpha) \end{pmatrix}. \tag{2.3}$$

With this notation in hand we present the following lemma.

[6] In some microeconometric datasets a substantial fraction of units never alter their value of X_t. For example, in Card (1996) few workers join or leave a union during the sample period.

Lemma 12. *Let $T = 2$. Suppose that Assumptions 4, 5 and 6 hold, and that θ is point-identified. Then, there exists $x_1 \in \{0,1\}$ and a non-zero function $\phi_{x_1} : \{0,1\}^3 \to \mathbb{R}$ such that:*

(i) for all $\alpha \in \mathcal{S}$ and $y_1 \in \{0,1\}$,

$$\sum_{y_2=0}^{1} \phi_{x_1}(y_1, y_2, 1) F(\theta + \alpha)^{y_2} [1 - F(\theta + \alpha)]^{1-y_2} = \sum_{y_2=0}^{1} \phi_{x_1}(y_1, y_2, 0) F(\alpha)^{y_2} [1 - F(\alpha)]^{1-y_2}; \tag{2.4}$$

(ii) for all $\alpha \in \mathcal{S}$ and $x_2 \in \{0,1\}$,

$$\begin{aligned} &\sum_{y_2=0}^{1} \sum_{y_1=0}^{1} \phi_{x_1}(y_1, y_2, x_2) \\ &\times F(\theta x_2 + \alpha)^{y_2} [1 - F(\theta x_2 + \alpha)]^{1-y_2} F(\theta x_1 + \alpha)^{y_1} [1 - F(\theta x_1 + \alpha)]^{1-y_1} = 0. \end{aligned} \tag{2.5}$$

The proof of Lemma 12 exploits the fact that, if θ is point-identified, then it is also locally point-identified. Together with the assumption that the parameter is regular, this allows us to apply a result of Bekker and Wansbeek (2001) regarding the identification of subvectors, which guarantees the existence of some $x_1 \in \{0,1\}$ such that $\nabla_{\theta'} Q_{x_1}$ does not belong to the range of the matrix $\begin{bmatrix} \nabla_{\pi'_{x_1}} Q_{x_1} & \nabla_{G'_{x_1}} Q_{x_1} \end{bmatrix}$. We then show, using (2.3), that this implies the existence of $\phi_{x_1} \neq 0$ such that (2.4) and (2.5) hold.

When the population parameter θ is point-identified, Lemma 12 suggests a method-of-moments approach to estimation. In such settings, $\phi_{X_{i1}}(Y_{i1}, Y_{i2}, X_{i2})$ will generally be a non-trivial function of θ. Let $\phi_{X_{i1}}(Y_{i1}, Y_{i2}, X_{i2}; \theta)$ be this function. Next, note that condition (2.4) in Lemma 12 corresponds to the conditional moment restriction

$$\mathbb{E}\left[\phi_{X_{i1}}(Y_{i1}, Y_{i2}, X_{i2}; \theta) \mid X_{i1}, X_{i2}, Y_{i1}, \alpha_i\right] = \mathbb{E}\left[\phi_{X_{i1}}(Y_{i1}, Y_{i2}, X_{i2}; \theta) \mid X_{i1}, Y_{i1}, \alpha_i\right], \tag{2.6}$$

while – continuing to maintain (2.4) – equation (2.5) implies the additional requirement that

$$\mathbb{E}\left[\phi_{X_{i1}}(Y_{i1}, Y_{i2}, X_{i2}; \theta) \mid X_{i1}, \alpha_i\right] = 0. \tag{2.7}$$

Analog estimators in point-identified models with feedback, based on these observations, are explored in our companion paper (Bonhomme et al., 2022).

This formulation clarifies that a necessary condition for point-identification of θ is the existence of a non-zero moment function, $\phi_{X_{i1}}(Y_{i1}, Y_{i2}, X_{i2}; \theta)$, with a mean that is invariant to X_{i2} given α_i and the past (i.e., the first period's covariate and outcome). Such a moment function is "feedback robust", in the sense that it remains valid across all possible feedback processes. This is the content of condition (2.4) in Lemma 12, while (2.5) imposes a similar invariance to the distribution of unobserved heterogeneity.

To show that point-identification fails, our focus here, we need to show that no such non-zero moment function exists. It turns out that there is a very simple condition for this in our model. Specifically, from Lemma 12 we obtain the following corollary.

Corollary 12.1. *Let $T = 2$. Suppose that Assumptions 4, 5 and 6 hold, and that 1, $F(\alpha)$, and $F(\theta + \alpha)$, for $\alpha \in \mathcal{S}$, are linearly independent, then θ is not point-identified.*

Corollary 12.1 shows that a necessary condition for identification of θ is that 1, $F(\alpha)$, and $F(\theta + \alpha)$, for $\alpha \in \mathcal{S}$, are linearly dependent. This condition arises directly from condition (2.4), which requires the existence of a moment function that is robust to unknown feedback. Indeed, one can show that 1, $F(\alpha)$, and $F(\theta + \alpha)$ are linearly dependent if and only if there exists a non-constant function ϕ such that

$$\mathbb{E}\left[\phi(Y_{it}, X_{it}) \,|\, X_{it}, \alpha_i\right] = \mathbb{E}\left[\phi(Y_{it}, X_{it}) \,|\, \alpha_i\right]. \tag{2.8}$$

However, the condition that 1, $F(\alpha)$, and $F(\theta + \alpha)$ be linearly dependent is restrictive, as we show in the next subsection.[7]

Remark 10. Despite the negative result of Corollary 12.1, the sign of θ is identified provided that Assumption 5 holds and $F(\cdot)$ is strictly increasing. Specifically, we show in Appendix 2.8.3 that

$$\text{sign}(\theta) = \text{sign}\left(\mathbb{E}\left[Y_{i2} - Y_{i1} \,|\, X_{i1} = 0\right]\right) = \text{sign}\left(\mathbb{E}\left[Y_{i1} - Y_{i2} \,|\, X_{i1} = 1\right]\right).$$

2.3.2 The logit model

Consider the logit model with a binary predetermined covariate, which corresponds to $F(u) = \frac{e^u}{1+e^u}$. In this case, the linear dependence condition of Corollary 12.1 requires that, for some non-zero triplet (A, B, C),

$$A\frac{e^{\theta+\alpha}}{1+e^{\theta+\alpha}} + B\frac{e^{\alpha}}{1+e^{\alpha}} + C = 0, \quad \text{for all } \alpha \in \mathcal{S}.$$

However, this implies

$$Ae^{\theta}e^{\alpha}(1+e^{\alpha}) + Be^{\alpha}(1+e^{\theta}e^{\alpha}) + C(1+e^{\alpha})(1+e^{\theta}e^{\alpha}) = 0, \quad \text{for all } \alpha \in \mathcal{S},$$

which is a quadratic polynomial equation in e^{α}. Therefore, provided that there are $K \geq 3$ values in $\mathcal{S}$, this implies

$$Ae^{\theta} + Be^{\theta} + Ce^{\theta} = 0, \quad Ae^{\theta} + B + (1+e^{\theta})C = 0, \quad C = 0,$$

which, provided that $\theta \neq 0$, entails

$$A = B = C = 0,$$

contradicting the assumption that (A, B, C) is non-zero.

We have thus proved the following corollary.

[7] While here we focus on a discrete $\mathcal{S}$ under Assumption 4, note that, when $\theta \neq 0$ and F is strictly increasing on $\mathbb{R}$, 1, $F(\alpha)$, and $F(\theta + \alpha)$, for $\alpha \in \mathbb{R}$, cannot be linearly dependent. If that were the case, then for some non-zero triplet (A, B, C) we would have $AF(\theta + \alpha) + BF(\alpha) + C = 0$ for all $\alpha \in \mathbb{R}$. This would imply, by taking $\alpha \to \pm\infty$ that $C = 0$ and $A + B = 0$, which would then imply $A = B = C = 0$ and contradict the assumption that (A, B, C) is non-zero.

Corollary 12.2. *Consider the logit model with $T = 2$. Suppose that Assumptions 4, 5 and 6 hold, that $\theta \neq 0$, and that $\mathcal{S}$ contains at least three points, then θ is not point-identified.*

A precedent to Corollary 12.2 is given in the unpublished working paper by Chamberlain (1993) mentioned in the introduction. In the model he considers, $X_{it} = Y_{i,t-1}$ is a lagged outcome, and $T = 2$ (hence, outcomes are observed for three periods). His model also includes an additional regressor: an indicator for period $t = 2$.

2.3.3 The exponential model

Suppose now that, for $u \geq 0$, $F(u) = 1 - e^{-u}$. This corresponds to the exponential binary choice model of Al-Sadoon et al. (2017). Note that here the support of $F(\cdot)$ is a strict subset of the real line. In this case, letting

$$A = e^{\theta},\ B = -1,\ C = 1 - e^{\theta},$$

we have

$$A[1 - e^{-(\theta+\alpha)}] + B[1 - e^{-\alpha}] + C = 0.$$

Hence the non point-identification condition of Corollary 12.1 is not satisfied in the exponential binary choice model.

In fact, in this case (2.4) and (2.5) are satisfied for

$$\phi_{x_1}(y_1, y_2, x_2; \theta) = (1 - y_2)e^{\theta x_2} - (1 - y_1)e^{\theta x_1},$$

and θ satisfies the conditional moment restriction

$$\mathbb{E}[\phi_{X_{i1}}(Y_{i1}, Y_{i2}, X_{i2}; \theta) \,|\, X_{i1}] = 0,$$

that is,

$$\mathbb{E}[(1 - Y_{i2})e^{\theta X_{i2}} - (1 - Y_{i1})e^{\theta X_{i1}} \,|\, X_{i1}] = 0. \tag{2.9}$$

See Wooldridge (1997) for several related results. Furthermore, one can show formally that θ is point-identified based on (2.9), see Appendix 2.8.4.

2.4 Failure of point-identification in T-period panels for $T > 2$

In this section we generalize our analysis to an arbitrary number of periods and state our main result.

2.4.1 Main result

The arguments laid out in the previous section extend to an arbitrary number of time periods, $T \geq 2$. Indeed, using a similar strategy to the proof of Lemma 12 and proceeding by induction, we obtain the following lemma.

Lemma 13. *Let $T \geq 2$. Suppose that Assumptions 4, 5 and 6 hold, and that θ is point-identified. Then, there exists $x_1 \in \{0,1\}$ and a non-zero function $\phi_{x_1} : \{0,1\}^{2T-1} \to \mathbb{R}$ such that:*

(i) for all $\alpha \in \mathcal{S}$, $s \in \{0, ..., T-2\}$, $y^{T-(s+1)} \in \{0,1\}^{T-(s+1)}$, $x^{T-(s+1)} \in \{0,1\}^{T-(s+1)}$,

$$\sum_{y^{T-s:T} \in \{0,1\}^{s+1}} \phi_{x_1}(y^T, x^{2:T}) \prod_{t=T-s}^{T} F(\theta x_t + \alpha)^{y_t} [1 - F(\theta x_t + \alpha)]^{1-y_t} \tag{2.10}$$

does not depend on $x^{T-s:T}$;

(ii) for all $\alpha \in \mathcal{S}$ and $x^{2:T} \in \{0,1\}^{T-1}$,

$$\sum_{y^T \in \{0,1\}^T} \phi_{x_1}(y^T, x^{2:T}) \prod_{t=1}^{T} F(\theta x_t + \alpha)^{y_t} [1 - F(\theta x_t + \alpha)]^{1-y_t} = 0. \tag{2.11}$$

Similarly to Lemma 12, Lemma 13 implies the existence of a moment function, with (generally) non-trivial dependence on θ, which is "feedback robust", in the sense that, for all $s \in \{0, ..., T-2\}$,

$$\mathbb{E}\left[\phi_{X_{i1}}(Y_i^T, X_i^{2:T}; \theta) \,|\, X_i^{T-s}, Y_i^{T-(s+1)}, \alpha_i\right] = \mathbb{E}\left[\phi_{X_{i1}}(Y_i^T, X_i^{2:T}; \theta) \,|\, X_i^{T-(s+1)}, Y_i^{T-(s+1)}, \alpha_i\right],$$

while also requiring that

$$\mathbb{E}\left[\phi_{X_{i1}}(Y_i^T, X_i^{2:T}; \theta) \,|\, X_{i1}, \alpha_i\right] = 0.$$

From Lemma 13 we obtain the following corollary, which we also prove by induction. This is our main result.

Corollary 13.1. *Let $T \geq 2$. Suppose that Assumptions 4, 5 and 6 hold, and that 1, $F(\alpha)$, and $F(\theta + \alpha)$, for $\alpha \in \mathcal{S}$, are linearly independent, then θ is not point-identified.*

2.4.2 Logit model

Using that, when $\theta \neq 0$, 1, $F(\alpha)$, and $F(\theta + \alpha)$, for $\alpha \in \mathcal{S}$, are linearly independent in the logit model, Corollary 13.1 implies that in the logit model with a binary predetermined covariate, θ is not point-identified irrespective of the number of time periods available.

Corollary 13.2. *Consider the logit model with $T \geq 2$. Suppose that Assumptions 4, 5 and 6 hold, that $\theta \neq 0$, and that $\mathcal{S}$ contains at least three points, then θ is not point-identified.*

This non point-identification result contrasts with prior work on logit panel data models. Under strict exogeneity, Rasch (1960) and Andersen (1970) have established that θ is point-identified under mild conditions on X_{it} whenever $T \geq 2$. In the dynamic logit model when $X_{it} = Y_{i,t-1}$, Chamberlain (1993) shows that θ is not point-identified when $T = 2$ (a result also obtained as an implication of Corollary 12.1). However, Chamberlain (1985a), and Honoré and Kyriazidou (2000) in a model with covariates, show that θ is point-identified under suitable conditions whenever $T \geq 3$.[8] By contrast, Corollary 13.2 shows that, when the feedback process through which current covariates are influenced by lagged outcomes is unrestricted, the failure of point-identification is pervasive irrespective of T, despite the logit structure.

2.5 Characterizing identified sets

The previous sections show that point-identification often fails in binary choice models with a predetermined covariate. In this section, we explore the degree of identification failure by presenting numerical calculations of the identified set Θ^I for specific parameter values. In the last part of the section we present calculations of the identified set for an average partial effect.

2.5.1 Linear programming representation

We show that the identified set Θ^I, defined by set (2.2) above, can be represented as a set of θ values for which a certain linear program has a solution. This characterization facilitates numerical computation of the identified set.

To present our construction, let us first focus on the $T = 2$ case, and suppose that Assumption 4 holds, so α_i has discrete support. For any hypothetical values $(\widetilde{\theta}, \widetilde{\pi}, \widetilde{G}) \in \Theta \times \Pi \times \mathcal{G}_2$, we define

$$\psi_{x_1}(x_2, y_1, \alpha) = \Pr\Big(X_{i2} = x_2, Y_{i1} = y_1, \alpha_i = \alpha \,|\, X_{i1} = x_1; \widetilde{\theta}, \widetilde{\pi}, \widetilde{G}\Big). \tag{2.12}$$

The right-hand-side of (2.12) is determined by the unknown heterogeneity distribution, the parametric likelihood for Y_1 (given X_1 and α), and the unknown feedback process for X_2. Finding Θ_I essentially involves repeatedly asking whether, for a given $\widetilde{\theta}$, there exists a valid feedback process and heterogeneity distributions consistent with the observed data distribution (and the parametric part of the model).

Specifically we first require that $\psi_{x_1}(x_2, y_1, \alpha)$ is a valid probability mass function:

$$\psi_{x_1}(x_2, y_1, \alpha) \geq 0, \quad \sum_{x_2=0}^{1} \sum_{y_1=0}^{1} \int_{\mathcal{S}} \psi_{x_1}(x_2, y_1, \alpha) d\mu(\alpha) = 1. \tag{2.13}$$

[8] Since in the dynamic logit model $X_{it} = Y_{i,t-1}$ is a lagged outcome, $T \geq 2$ (respectively, $T \geq 3$) requires that individual outcomes be available for at least three (resp., four) periods.

Second, we check that it is consistent with the parametric likelihood model for Y_1 given X_1 and α:

$$\sum_{x_2=0}^{1} \psi_{x_1}(x_2, y_1, \alpha) = F(\widetilde{\theta} x_1 + \alpha)^{y_1} [1 - F(\widetilde{\theta} x_1 + \alpha)]^{1-y_1} \sum_{x_2=0}^{1} \sum_{y_1=0}^{1} \psi_{x_1}(x_2, y_1, \alpha). \tag{2.14}$$

Finally, we conclude that $\widetilde{\theta} \in \Theta^I$ if and only if

$$Q_{x_1}(y_2, y_1, x_2; \theta, \pi, G) = \int_{\mathcal{S}} F(\widetilde{\theta} x_2 + \alpha)^{y_2} [1 - F(\widetilde{\theta} x_2 + \alpha)]^{1-y_2} \psi_{x_1}(x_2, y_1, \alpha) d\mu(\alpha), \tag{2.15}$$

for some vectors ψ_{x_1} also satisfying (2.13) and (2.14) for $x_1 \in \{0, 1\}$. Condition (2.15) ensures compatibility with the likelihood contribution for the period 2 outcome, Y_2.

Since all of the equalities and inequalities in (2.13), (2.14) and (2.15) are linear in ψ_{x_1}, it follows that one can verify whether $\widetilde{\theta} \in \Theta^I$ by checking the existence of a solution to a finite-dimensional linear program.[9] We provide details about computation in Appendix 2.8.8.

The characterization of Θ^I in (2.13), (2.14) and (2.15) remains valid when Assumption 4 does not hold, and α_i has continuous support. In that case, one needs to interpret ψ_{x_1} in (2.12) as the product between the density of α_i conditional on (X_{i2}, Y_{i1}) and the probability of (X_{i2}, Y_{i1}), both of them conditional on X_{i1} and for hypothetical parameter values. The resulting linear program is infinite-dimensional in that case.

The linear programming representation of Θ^I extends to any number $T \geq 2$ of periods. To see this, let, for some $(\widetilde{\theta}, \widetilde{\pi}, \widetilde{G}) \in \Theta \times \Pi \times \mathcal{G}_T$,

$$\psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) = \Pr\Big(X_i^{2:T} = x^{2:T}, Y_i^{T-1} = y^{T-1}, \alpha_i = \alpha \,|\, X_{i1} = x_1; \widetilde{\theta}, \widetilde{\pi}, \widetilde{G}\Big),$$

with a similar definition when the support of α_i is not discrete and Assumption 4 does not hold. In Appendix 2.8.7 we derive the following characterization of the (sharp) identified set Θ^I.

Proposition 5. *(Identified Set) $\widetilde{\theta} \in \Theta^I$ if, and only if,*

$$Q_{x_1}(y^T, x^{2:T}; \theta, \pi, G) = \int_{\mathcal{S}} F(\widetilde{\theta} x_T + \alpha)^{y_T} [1 - F(\widetilde{\theta} x_T + \alpha)]^{1-y_T} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) d\mu(\alpha), \tag{2.16}$$

[9] Note that, to compute the identified set under the assumption of strict exogeneity, one can simply modify this approach by adding to (2.13), (2.14) and (2.15) the additional restriction

$$\frac{\psi_{x_1}(x_2, 1, \alpha)}{F(\widetilde{\theta} x_1 + \alpha)} = \frac{\psi_{x_1}(x_2, 0, \alpha)}{1 - F(\widetilde{\theta} x_1 + \alpha)} \quad \text{for all } (x_2, x_1, \alpha),$$

which is also linear in ψ_{x_1}. The fact that, under strict exogeneity, Θ^I can be computed using linear programming was first established by Honoré and Tamer (2006).

for some integrable functions $\psi_{x_1} : \{0,1\}^{2T-2} \times \mathcal{S} \to \mathbb{R}$, $x_1 \in \{0,1\}$, *satisfying*

$$\psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) \geq 0, \quad \sum_{x^{2:T} \in \{0,1\}^{T-1}} \sum_{y^{T-1} \in \{0,1\}^{T-1}} \int_{\mathcal{S}} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) d\mu(\alpha) = 1, \tag{2.17}$$

and, for all $s \in \{2, ..., T\}$,[10] *also satisfying*

$$\begin{aligned} &\sum_{x^{s:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s:T-1} \in \{0,1\}^{T-s}} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) \\ &= F(\widetilde{\theta} x_{s-1} + \alpha)^{y_{s-1}} [1 - F(\widetilde{\theta} x_{s-1} + \alpha)]^{1-y_{s-1}} \sum_{x^{s:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s-1:T-1} \in \{0,1\}^{T-s+1}} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha). \end{aligned} \tag{2.18}$$

Proposition 5 shows that one can verify whether $\widetilde{\theta} \in \Theta^I$ by checking the feasibility of a (finite- or infinite-dimensional) linear program. In a setting with lagged outcomes and strictly exogenous covariates, Honoré and Tamer (2006) provided an analogous linear programming representation of the identified set. By contrast, in Proposition 5 we characterize the identified set of θ in the general predetermined case where the Granger condition fails; i.e., when $G_{y^{t-1},x^{t-1}}(\alpha)$ may depend on y^{t-1}, a situation that Honoré and Tamer (2006) did not consider but anticipated in their conclusion.

2.5.2 Numerical illustration

In this section we compute identified sets Θ^I in logit and probit models for a set of example data generating processes (DGPs). In the DGPs, X_{it} follows a Bernoulli distribution on $\{0,1\}$ with probabilities $(\frac{1}{2}, \frac{1}{2})$, independent over time, and α_i takes $K = 31$ values with probabilities closely resembling those of a standard normal (a specification we borrow from Honoré and Tamer, 2006), and is drawn independently of $(X_{i1}, ..., X_{iT})$. In the logit case, $F(u) = \frac{e^u}{1+e^u}$, and in the probit case, $F(u) = \Phi(u)$ for Φ the standard normal cdf. Lastly, we vary θ between -1 and 1. Note that X_{it} is strictly exogenous in this data generating process. We characterize identified sets in two scenarios: assuming that X_{it} are strictly exogenous, and only assuming that X_{it} are predetermined.

In Figure 2.1 we report our numerical calculations of the identified set Θ^I for the logit model (in the left column panels) and for the probit model (in the right column panels). The three vertical panels correspond to the $T = 2, 3, 4$ cases, respectively. In each graph, we report two sets of upper and lower bounds: those computed while maintaining the strict

[10] For $s = T$, restriction (2.18) should be read as

$$\sum_{x_T=0}^{1} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) = F(\widetilde{\theta} x_{T-1} + \alpha)^{y_{T-1}} [1 - F(\widetilde{\theta} x_{T-1} + \alpha)]^{1-y_{T-1}} \sum_{x_T=0}^{1} \sum_{y_{T-1}=0}^{1} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha).$$

exogeneity assumption (in dashed lines) and those computed maintaining just predeterminedness (in solid lines). We report the true parameter θ on the x-axis. To compute the sets, we assume that α_i has the same $K = 31$ points of support as in the DGP. We also experimented with fewer and additional support points, as we report below.

Focusing first on the logit case, shown in the left column of Figure 2.1, we see that the identified set Θ^I under strict exogeneity is a singleton for any value of θ and irrespective of T. This is not surprising since θ is point-identified in the static logit model. In contrast, the upper and lower bounds of the identified set do not coincide in the predetermined case, consistent with our non point-identification result. At the same time, the identified sets appear rather narrow, even when $T = 2$, and the width of the set tends to decrease rapidly when T increases to three and four periods. This is qualitatively similar to the observation of Honoré and Tamer (2006), who focused on dynamic probit models and found that the width of the identified set tends to decrease rapidly with T.

Focusing next on the probit case, shown in the right column of Figure 2.1, we see that the identified set Θ^I under strict exogeneity is not a singleton. Moreover, allowing the covariate to be predetermined increases the width of the identified set. However, as in the logit case, the sets appear rather narrow, even when $T = 2$, and their widths decrease quickly as T increases. Of course, these observations are specific to a particular data-generating process and the corresponding bounds may be wide for other DGPs.

The results in Figure 2.1 are obtained by assuming that the researcher knows the (finite) support of α_i. This approach is similar to the one in Honoré and Tamer (2006). Alternatively, one may wish to characterize the identified set in a class of models where α_i is continuous, e.g., when $\mathcal{S} = \mathbb{R}$ and μ is the Lebesgue measure. Doing so, as noted earlier, requires approximating an infinite-dimensional linear program. In Appendix Figure 2.3, we go take a heuristic step in this direction by reporting numerical approximations to the identified sets, for $T = 2$, obtained by taking $K = 5$, $K = 50$, and $K = 500$ points of support for α_i, respectively, where the points of support are equidistant percentiles of a standard normal distribution. We find very minor differences compared to the case $K = 31$ that we report in Figure 2.1. While we do not provide a formal analysis of numerical approximation properties, this suggests that identified sets under continuous α_i may not be markedly different from the ones in Figure 2.1.

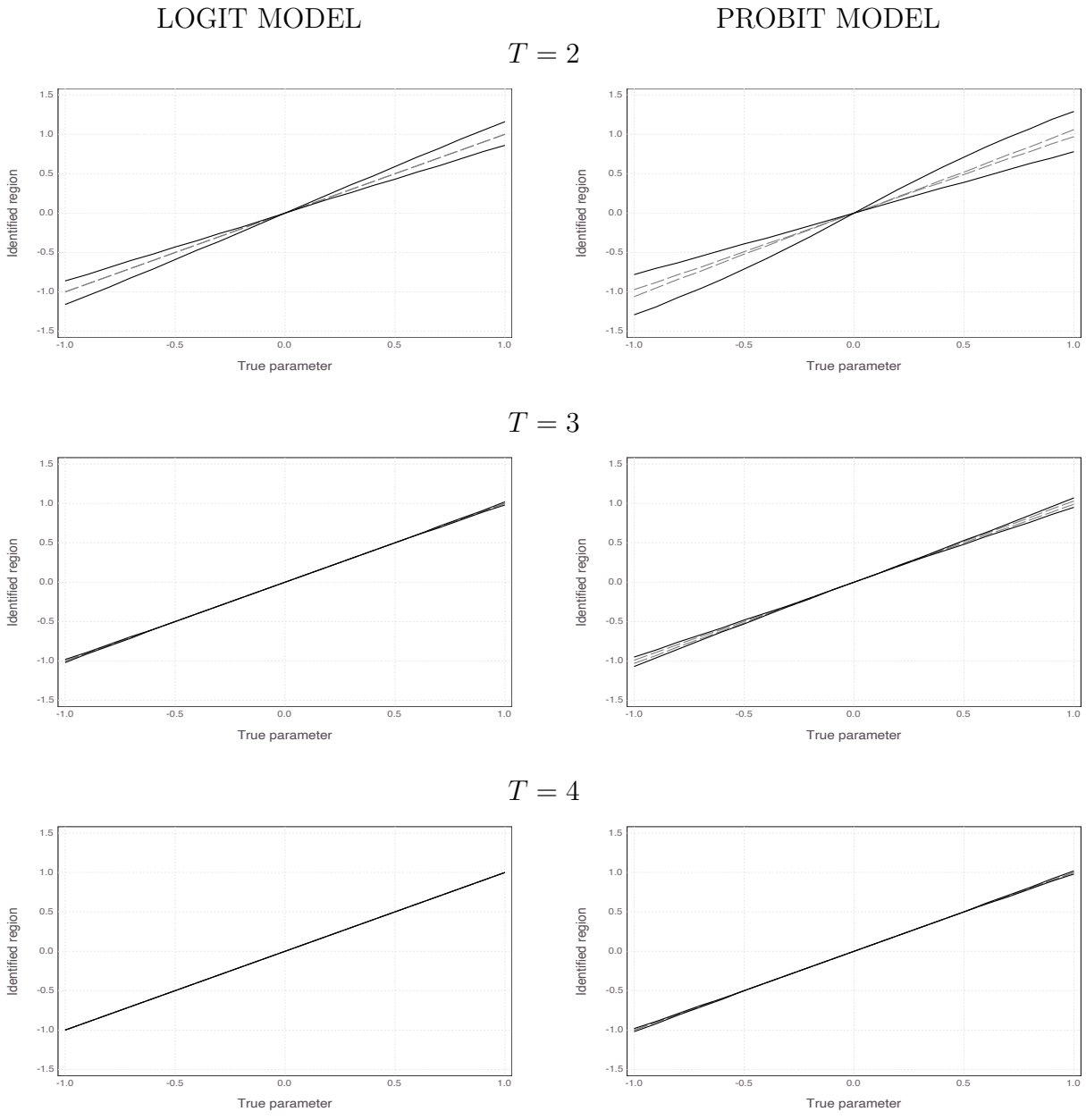
Overall, these calculations suggest that, while relaxing strict exogeneity tends to increase the widths of the bounds, the identified sets under predeterminedness can be informative even when the number of periods is very small. To reiterate, these conclusions are based on a particular set of example DGPs.

2.5.3 Average partial effect

Although our focus in this chapter is on the parameter θ, in applications researchers are often interested in average partial effects such as

$$\Delta = \mathbb{E}[\Pr\big(Y_{it} = 1 \,|\, X_{it} = 1, \alpha_i\big) - \Pr\big(Y_{it} = 1 \,|\, X_{it} = 0, \alpha_i\big)], \tag{2.19}$$

Figure 2.1: Identified sets in logit and probit models



Notes: Upper and lower bounds of the identified set Θ^I in a logit model (left column) and a probit model (right column), for $T = 2, 3, 4$. The identified sets under strict exogeneity are indicated by the dashed lines, the sets under predeterminedness are indicated by the solid lines. The population value of θ is given on the x-axis.

where the expectation is taken with respect to the distribution of α_i.

The identified set for Δ can also be characterized as the solution to a linear program. Indeed, it follows from Proposition 5 that $\widetilde{\Delta}$ is in the identified set of Δ if and only if there exists $\widetilde{\theta}$, ψ_0 and ψ_1 such that (2.16), (2.17), and (2.18) hold, and

$$\widetilde{\Delta} = \int_{\mathcal{S}} [F(\widetilde{\theta} + \alpha) - F(\alpha)] \sum_{x_1 \in \{0,1\}} q_{x_1} \sum_{x^{2:T} \in \{0,1\}^{T-1}} \sum_{y^{T-1} \in \{0,1\}^{T-1}} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) d\mu(\alpha), \quad (2.20)$$

where $q_{x_1} = \Pr(X_{i1} = x_1)$. For any given $\widetilde{\theta} \in \Theta^I$, we can therefore compute the set of $\widetilde{\Delta}$ parameters in the identified set by solving a linear program. We provide details about computation in Appendix 2.8.8.

In Figure 2.2 we report our computations of the identified set for the average partial effect Δ, relying on the same parameter values and DGP as before. Focusing first on the logit case, shown in the left column of the figure, we see that the identified set under strict exogeneity is not a singleton, except when the true θ and Δ are equal to zero. This is not surprising, since average partial effects generally fail to be point-identified in binary choice models, even when covariates are strictly exogenous. Yet, the sets seem rather narrow, even when $T = 2$. Allowing the covariate to be predetermined increases the widths of the sets, however the increase is relatively moderate. Moreover, the sets under predeterminedness are very tight whenever $T \geq 3$.

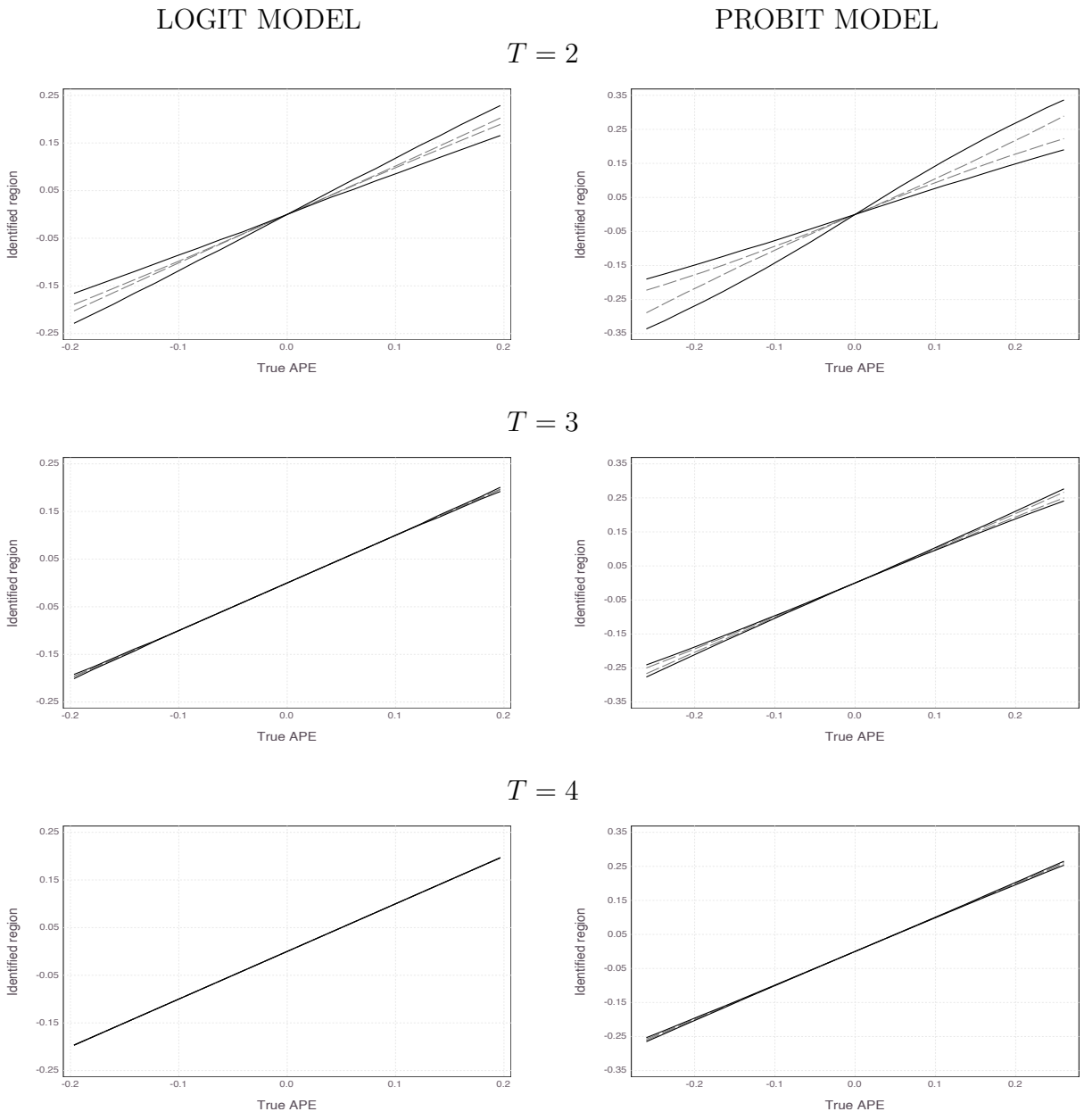
Focusing next on the probit case, shown in the right column of Figure 2.2, we see that although the sets appear wider than in the logit case, relaxing strict exogeneity only moderately increases the widths of the sets, especially when $T \geq 3$.

Lastly, while we compute the sets in Figure 2.2 under the assumption that α_i has the same $K = 31$ points of support as in the DGP, in Appendix Figure 2.4 we report approximations of the sets, for $T = 2$, obtained using $K = 5$, $K = 50$, and $K = 500$ points of support for α_i. The sets appear very similar to the ones based on $K = 31$ points of support shown in Figure 2.2. However, in this case as well, we do not formally analyze the numerical approximation of the identified sets under continuous α_i.

2.6 Restrictions on the feedback process

Our analysis suggests that failures of point-identification are commonplace in binary choice models with a predetermined covariate. In this section we describe possible restrictions on the model that can strengthen its identification content. We focus on restrictions on the feedback process, since restrictions on individual heterogeneity are rarely motivated by the economic context.

Figure 2.2: Identified sets for average partial effects in logit and probit models



Notes: Upper and lower bounds of the identified set for the average partial effect in a logit model (left column) and a probit model (right column), for $T = 2, 3, 4$. The identified sets under strict exogeneity are indicated by the dashed lines, the sets under predeterminedness are indicated by the solid lines. The population value of the average partial effect is given on the x-axis.

2.6.1 Homogeneous feedback

In some applications one may want to restrict the feedback process to not depend on time-invariant heterogeneity; that is, to impose that

$$\Pr\left(X_{it} = 1 \,|\, Y_i^{t-1} = y^{t-1}, X_i^{t-1} = x^{t-1}, \alpha_i = \alpha\right) = G^t_{y^{t-1},x^{t-1}} \tag{2.21}$$

is independent of α. For example, in structural dynamic discrete choice models, researchers may be willing to model the law of motion of state variables such as dynamic production inputs as homogeneous across units. Kasahara and Shimotsu (2009) show how this assumption can help identification in these models. Here we study how a homogeneity assumption can lead to tighter identified sets in our setting.

To proceed, we focus on the case where $T = 2$. Given (2.21), the likelihood function takes the form

$$\begin{aligned}
&\Pr\left(Y_{i2} = y_2, X_{i2} = x_2, Y_{i1} = y_1 \,|\, X_{i1} = x_1\right) \\
&= \left\{ \int_{\mathcal{S}} F(\theta x_2 + \alpha)^{y_2} [1 - F(\theta x_2 + \alpha)]^{1-y_2} F(\theta x_1 + \alpha)^{y_1} [1 - F(\theta x_1 + \alpha)]^{1-y_1} \pi_{x_1}(\alpha) d\mu(\alpha) \right\} \\
&\qquad \times [G^2_{y_1,x_1}]^{x_2} [1 - G^2_{y_1,x_1}]^{1-x_2},
\end{aligned}$$

where the likelihood factors due to the fact that the feedback process does not depend on α. Hence, under Assumption 5 (which avoids division by zero) we have

$$\begin{aligned}
&\frac{\Pr\left(Y_{i2} = y_2, X_{i2} = x_2, Y_{i1} = y_1 \,|\, X_{i1} = x_1\right)}{[G^2_{y_1,x_1}]^{x_2} [1 - G^2_{y_1,x_1}]^{1-x_2}} \\
&= \int_{\mathcal{S}} F(\theta x_2 + \alpha)^{y_2} [1 - F(\theta x_2 + \alpha)]^{1-y_2} F(\theta x_1 + \alpha)^{y_1} [1 - F(\theta x_1 + \alpha)]^{1-y_1} \pi_{x_1}(\alpha) d\mu(\alpha).
\end{aligned} \tag{2.22}$$

A key observation to make about (2.22) is its right-hand-side coincides with the likelihood function of a binary choice model with a strictly exogenous covariate (where in addition α_i is independent of X_{i2} given X_{i1}). In turn, the left-hand side is weighted by the inverse of the feedback process. This is similar to the inverse-probability-of-treatment-weighting approach to dynamic treatment effect analysis in Jamie Robins' work (e.g., Robins, 2000), with the difference that here we focus on panel data models with fixed effects.

The similarity between (2.22) and the strictly exogenous case directly delivers point-identification results and consistent estimators. For example, suppose that F is logistic. Given that the left-hand side of (2.22) is point-identified, it follows from standard arguments (Rasch, 1960, Andersen, 1970) that θ is point-identified. Moreover, a consistent estimator of θ is obtained by maximizing the weighted conditional logit log-likelihood

$$\sum_{i=1}^{n} \widehat{\omega}_i \mathbf{1}\{Y_{i1} + Y_{i2} = 1\}$$
$$\times \left\{ Y_{i1} \ln \left(\frac{\exp\left(\widetilde{\theta} X_{i1}\right)}{\exp\left(\widetilde{\theta} X_{i1}\right) + \exp\left(\widetilde{\theta} X_{i2}\right)} \right) + Y_{i2} \ln \left(\frac{\exp\left(\widetilde{\theta} X_{i2}\right)}{\exp\left(\widetilde{\theta} X_{i1}\right) + \exp\left(\widetilde{\theta} X_{i2}\right)} \right) \right\},$$

with weights

$$\widehat{\omega}_i = \left\{ [\widehat{G}^2_{Y_{i1},X_{i1}}]^{X_{i2}} [1 - \widehat{G}^2_{Y_{i1},X_{i1}}]^{1-X_{i2}} \right\}^{-1},$$

for $\widehat{G}^2_{y_1,x_1}$ a consistent estimate of the homogeneous feedback probabilities.[11]

2.6.2 Markovian feedback

Another possible restriction on the feedback process is a Markovian condition, such as

$$\Pr\left(X_{it} = 1 \,|\, Y_i^{t-1} = y^{t-1}, X_i^{t-1} = x^{t-1}, \alpha_i = \alpha\right) = G^t_{y_{t-1},x_{t-1}}(\alpha) \tag{2.23}$$

is independent of (y^{t-2}, x^{t-2}). Such a condition may be natural in models where X_{it} is the state variable in the agent's economic problem (as in Rust, 1987 and Kasahara and Shimotsu, 2009, for example).

In order to characterize the identified set Θ^I with the Markovian condition (2.23) added, we augment the restrictions (2.16), (2.17) and (2.18) with the fact that, for all $s \in \{2, ..., T\}$,

$$\frac{\sum_{x^{s+1:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s:T-1} \in \{0,1\}^{T-s}} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha)}{\sum_{x^{s:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s:T-1} \in \{0,1\}^{T-s}} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha)}$$

does not depend on (y^{s-2}, x^{s-2}).[12]

A difficulty arises in this case since this additional set of restrictions is not linear in ψ_{x_1}. As a result, one would need to use different techniques to characterize the identified set in the spirit of Proposition 5, and to establish conditions for (the failure of) point-identification in the spirit of Corollary 13.1. Given this, we leave the analysis of identification in models with Markovian feedback processes to future work.

[11] The analysis in this subsection is not restricted to the binary covariate case. However, when X_{it} are continuous, demonstrating $\sqrt{n}$ consistency of $\widehat{\theta}$ would generally require imposing rate-of-convergence and other requirements on the first-step estimation of the $\widehat{\omega}_i$ weights.

[12] When $s = T$, this requires that $\frac{\psi_{x_1}(x^{2:T}, y^{T-1}, \alpha)}{\sum_{x_T=0}^{1} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha)}$ does not depend on (y^{T-2}, x^{T-2}).

2.7 Conclusion

In this chapter we study a binary choice model with a binary predetermined covariate. We find that failures of point-identification are widespread in this setting. Point-identification fails in many binary choice models, with apparently only a few exceptions (such as the exponential model). At the same time, our numerical calculations of identified sets suggest that the bounds on the parameter can be narrow, even in very short panels. This suggests that, while the strict exogeneity assumption has identifying content, models with predetermined covariates and feedback may still lead to informative empirical conclusions, both for the coefficients of the covariates and for average partial effects.

Our analysis of models with a binary covariates can easily be extended to handle general discrete covariates with finite support. In particular, for θ to be regularly point-identified there need to exist $x_1 \neq x_2$ in the support of X_{it} such that 1, $F(\theta' x_1 + \alpha)$, and $F(\theta' x_2 + \alpha)$, for $\alpha \in \mathcal{S}$, are linearly dependent. This condition fails in many popular specifications such as the logit. In turn, when X_{it} has finite, non-binary support, the identified set can still be computed as a solution to a linear program, analogously to Proposition 5. However, the extension to continuous covariates is not straightforward in our setting, in particular since the notion of regularity maintained by Assumption 6 no longer applies.

Finally, although we have analyzed a binary choice model, our techniques can be used to study other models with stronger identification content, such as models for count data (e.g., Poisson regression, Wooldridge, 1997, Blundell et al., 2002) and models with continuous outcomes (e.g., censored regression, Honoré and Hu, 2004, and duration models, Chamberlain, 1985a). Deriving sequential moment restrictions in such nonlinear models was considered by Chamberlain (2022) and is further explored in our companion paper (Bonhomme et al., 2022).

2.8 Appendix: proofs and additional materials

2.8.1 Proof of Lemma 12

For any $m \times n$ matrix A, we will denote as

$$\mathcal{R}(A) = \{Au \,:\, u \in \mathbb{R}^n\}$$

the range of A,

$$\mathcal{N}(A) = \{u \in \mathbb{R}^n \,:\, Au = 0\}$$

the null space of A, and $A^\dagger$ the Moore-Penrose generalized inverse of A.

We now proceed to prove Lemma 12. Since θ is point-identified, it is locally point-identified. Since (θ, π, G) is a regular point of $\nabla Q(\theta, \pi, G)$ by Assumption 6, it follows from Theorem

8 in Bekker and Wansbeek (2001) that

$$\nabla_{\theta'} Q \notin \mathcal{R}\left(\begin{bmatrix} \nabla_{\pi_1'} Q_1 & \nabla_{G_1'} Q_1 & 0 & 0 \\ 0 & 0 & \nabla_{\pi_0'} Q_0 & \nabla_{G_0'} Q_0 \end{bmatrix}\right). \tag{2.24}$$

Therefore, there must exist $x_1 \in \{0,1\}$ such that

$$\nabla_{\theta'} Q_{x_1} \notin \mathcal{R}\left(\begin{bmatrix} \nabla_{\pi_{x_1}'} Q_{x_1} & \nabla_{G_{x_1}'} Q_{x_1} \end{bmatrix}\right), \tag{2.25}$$

and in the rest of the proof we will fix this x_1 value.
Let $\widetilde{\phi}_{x_1}$ denote the projection of $\nabla_{\theta'} Q_{x_1}$ onto the orthogonal complement of the vector space spanned by the columns of $\begin{bmatrix} \nabla_{\pi_{x_1}'} Q_{x_1} & \nabla_{G_{x_1}'} Q_{x_1} \end{bmatrix}$; that is,

$$\widetilde{\phi}_{x_1} = \nabla_{\theta'} Q_{x_1} - \begin{bmatrix} \nabla_{\pi_{x_1}'} Q_{x_1} & \nabla_{G_{x_1}'} Q_{x_1} \end{bmatrix} \begin{bmatrix} \nabla_{\pi_{x_1}'} Q_{x_1} & \nabla_{G_{x_1}'} Q_{x_1} \end{bmatrix}^{\dagger} \nabla_{\theta'} Q_{x_1}.$$

It follows from (2.25) that $\widetilde{\phi}_{x_1} \neq 0$. Moreover, since $\iota' Q_{x_1}(\theta, \pi, G) = 1$, where ι denotes a conformable vector of ones, we have

$$\iota' \nabla_{\theta'} Q_{x_1} = 0, \;\; \iota' \nabla_{\pi_{x_1}'} Q_{x_1} = 0, \;\; \iota' \nabla_{G_{x_1}'} Q_{x_1} = 0. \tag{2.26}$$

It follows that $\iota' \widetilde{\phi}_{x_1} = 0$, implying that $\widetilde{\phi}_{x_1}$ cannot be constant.
Now, since $v' \widetilde{\phi}_{x_1} = 0$ for all $v \in \mathcal{R}\left(\begin{bmatrix} \nabla_{\pi_{x_1}'} Q_{x_1} & \nabla_{G_{x_1}'} Q_{x_1} \end{bmatrix}\right)$, we have

$$\widetilde{\phi}_{x_1} \in \mathcal{N}(\nabla_{\pi_{x_1}} Q_{x_1}') \cap \mathcal{N}(\nabla_{G_{x_1}} Q_{x_1}').$$

Next, let $P_\theta(x_1, \alpha)$ be the 8×1 vector with elements

$$\Pr\big(Y_{i2} = y_2, X_{i2} = x_2, Y_{i1} = y_1 \,|\, X_{i1} = x_1, \alpha_i = \alpha\big),$$

for $(y_2, x_2, y_1) \in \{0,1\}^3$. Since $\widetilde{\phi}_{x_1} \in \mathcal{N}(\nabla_{\pi_{x_1}} Q_{x_1}')$, we have, for all $\alpha \in \mathcal{S}$,

$$\widetilde{\phi}_{x_1}' P_\theta(x_1, \alpha) = \widetilde{\phi}_{x_1}' P_\theta(x_1, \underline{\alpha}_K) \equiv C_{x_1},$$

where we have used the fact that $\pi_{x_1}(\underline{\alpha}_K) = 1 - \sum_{k=1}^{K-1} \pi_{x_1}(\underline{\alpha}_k)$.
Let us define the following demeaned version of $\widetilde{\phi}_{x_1}$:[13]

$$\phi_{x_1} = \widetilde{\phi}_{x_1} - C_{x_1} \iota.$$

[13]The 8×1 vector ϕ_{x_1} represents a function $\phi_{x_1} : \{0,1\}^3 \mapsto \mathbb{R}$. With some abuse of terminology we sometimes refer to ϕ_{x_1} as a vector and sometimes as a function.

Note that, since $\widetilde{\phi}_{x_1}$ is not constant, it follows that $\phi_{x_1} \neq 0$. Moreover, using (2.25) and (2.26) we have

$$\phi_{x_1} \in \mathcal{N}(\nabla_{\pi_{x_1}} Q'_{x_1}) \cap \mathcal{N}(\nabla_{G_{x_1}} Q'_{x_1}),$$

from which it follows that

$$\text{(i)} \quad \nabla_{\pi_{x_1}} Q'_{x_1} \phi_{x_1} = 0, \quad \text{(ii)} \quad \nabla_{G_{x_1}} Q'_{x_1} \phi_{x_1} = 0.$$

We are now going to use (i) and (ii) to show (2.4)-(2.5). From (ii) we get, for all $\alpha \in \mathcal{S}$,

$$\begin{aligned}
&\pi_{x_1}(\alpha)\Big(\phi_{x_1}(1,1,1)F(\theta+\alpha)F(\theta x_1+\alpha) - \phi_{x_1}(1,1,0)F(\alpha)F(\theta x_1+\alpha) \\
&+ \phi_{x_1}(1,0,1)[1-F(\theta+\alpha)]F(\theta x_1+\alpha) - \phi_{x_1}(1,0,0)[1-F(\alpha)]F(\theta x_1+\alpha)\Big) = 0, \\
&\pi_{x_1}(\alpha)\Big(\phi_{x_1}(0,1,1)F(\theta+\alpha)[1-F(\theta x_1+\alpha)] - \phi_{x_1}(0,1,0)F(\alpha)[1-F(\theta x_1+\alpha)] \\
&+ \phi_{x_1}(0,0,1)[1-F(\theta+\alpha)][1-F(\theta x_1+\alpha)] - \phi_{x_1}(0,0,0)[1-F(\alpha)][1-F(\theta x_1+\alpha)]\Big) = 0.
\end{aligned}$$

This implies, using Assumption 5,

$$\begin{aligned}
&\phi_{x_1}(1,1,1)F(\theta+\alpha) - \phi_{x_1}(1,1,0)F(\alpha) \\
&+ \phi_{x_1}(1,0,1)[1-F(\theta+\alpha)] - \phi_{x_1}(1,0,0)[1-F(\alpha)] = 0, \\
&\phi_{x_1}(0,1,1)F(\theta+\alpha) - \phi_{x_1}(0,1,0)F(\alpha) \\
&+ \phi_{x_1}(0,0,1)[1-F(\theta+\alpha)] - \phi_{x_1}(0,0,0)[1-F(\alpha)] = 0,
\end{aligned}$$

which coincides with (2.4).
Lastly, from (i) we get, for all $\alpha \in \mathcal{S}$,

$$\begin{aligned}
\phi'_{x_1} P_\theta(x_1,\alpha) &= \phi'_{x_1} P_\theta(x_1,\underline{\alpha}_K) \\
&= \widetilde{\phi}'_{x_1} P_\theta(x_1,\underline{\alpha}_K) - C_{x_1} \underbrace{\iota' P_\theta(x_1,\underline{\alpha}_K)}_{=1} \\
&= \widetilde{\phi}'_{x_1} P_\theta(x_1,\underline{\alpha}_K) - \widetilde{\phi}'_{x_1} P_\theta(x_1,\underline{\alpha}_K) \\
&= 0,
\end{aligned}$$

which can be equivalently written as

$$\sum_{y_2=0}^{1}\sum_{x_2=0}^{1}\sum_{y_1=0}^{1} \phi_{x_1}(y_1,y_2,x_2)\Pr\big(Y_{i2}=y_2, X_{i2}=x_2, Y_{i1}=y_1 \,|\, X_{i1}=x_1, \alpha_i=\alpha;\theta\big) = 0.$$

Now, using (2.4), this implies that, for all $x_2 \in \{0,1\}$,

$$\sum_{y_2=0}^{1}\sum_{y_1=0}^{1}\phi_{x_1}(y_1,y_2,x_2)$$
$$\times \Pr\big(Y_{i2}=y_2 \,|\, X_{i2}=x_2,\alpha_i=\alpha;\theta\big)\Pr\big(Y_{i1}=y_1 \,|\, X_{i1}=x_1,\alpha_i=\alpha;\theta\big)=0,$$

which coincides with (2.5).

2.8.2 Proof of Corollary 12.1

The proof is by contradiction. Suppose that θ is point-identified. Then by (2.4) we have, for some $x_1 \in \{0,1\}$, and for all $y_1 \in \{0,1\}$ and $\alpha \in \mathcal{S}$,

$$\begin{aligned}&\phi_{x_1}(y_1,0,1)[1-F(\theta+\alpha)]+\phi_{x_1}(y_1,1,1)F(\theta+\alpha)\\&=\phi_{x_1}(y_1,0,0)[1-F(\alpha)]+\phi_{x_1}(y_1,1,0)F(\alpha).\end{aligned}$$

Since 1, $F(\alpha)$, and $F(\theta+\alpha)$, for $\alpha \in \mathcal{S}$, are linearly independent, we thus have, for all $y_1 \in \{0,1\}$,

$$\phi_{x_1}(y_1,0,1)=\phi_{x_1}(y_1,1,1)=\phi_{x_1}(y_1,0,0)=\phi_{x_1}(y_1,1,0). \tag{2.27}$$

Next, using (2.5) at $x_2=1$ we have

$$\begin{aligned}&\phi_{x_1}(1,1,1)F(\theta+\alpha)F(\theta x_1+\alpha)+\phi_{x_1}(0,1,1)F(\theta+\alpha)[1-F(\theta x_1+\alpha)]\\&+\phi_{x_1}(1,0,1)[1-F(\theta+\alpha)]F(\theta x_1+\alpha)+\phi_{x_1}(0,0,1)[1-F(\theta+\alpha)][1-F(\theta x_1+\alpha)]=0.\end{aligned}$$

Using (2.27) then gives

$$\phi_{x_1}(1,1,1)F(\theta x_1+\alpha)+\phi_{x_1}(0,1,1)[1-F(\theta x_1+\alpha)]=0.$$

Now, since 1 and $F(\theta x_1+\alpha)$, for $\alpha \in \mathcal{S}$, are linearly independent, it follows that

$$\phi_{x_1}(1,1,1)=\phi_{x_1}(0,1,1)=0.$$

Using (2.27) then also gives

$$\phi_{x_1}(1,0,1)=\phi_{x_1}(0,0,1)=0.$$

Lastly, repeating the same argument starting with (2.5) at $x_2=0$ gives

$$\phi_{x_1}(1,1,0)=\phi_{x_1}(0,1,0)=\phi_{x_1}(1,0,0)=\phi_{x_1}(0,0,0)=0.$$

It follows that $\phi_{x_1}=0$, which leads to a contradiction.

2.8.3 Proof of remark 10 (sign identification of θ)

Note that

$$\begin{aligned}
&\mathbb{E}\left[Y_{i2} - Y_{i1} \,|\, X_{i1} = 0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_{i2} \,|\, X_{i2}, Y_{i1}, X_{i1} = 0, \alpha_i\right] - \mathbb{E}\left[Y_{i1} \,|\, X_{i1} = 0, \alpha_i\right] \,|\, X_{i1} = 0\right] \\
&= \mathbb{E}\left[F(\theta X_{i2} + \alpha_i) - F(\alpha_i) \,|\, X_{i1} = 0\right] \\
&= \mathbb{E}\left[(F(\theta + \alpha_i) - F(\alpha_i))X_{i2}Y_{i1} + (F(\theta + \alpha_i) - F(\alpha_i))X_{i2}(1 - Y_{i1}) \,|\, X_{i1} = 0\right] \\
&= \int_{\mathcal{S}} \sum_{y_1=0}^{1} (F(\theta + \alpha) - F(\alpha)) \underbrace{G^2_{y_1,0}(\alpha)F(\alpha)^{y_1}(1 - F(\alpha))^{1-y_1}\pi_0(\alpha)}_{>0 \text{ by Assumption 5}} d\mu(\alpha).
\end{aligned} \tag{2.28}$$

If $\theta = 0$, (2.28) implies that $\mathbb{E}\left[Y_{i2} - Y_{i1} \,|\, X_{i1} = 0\right] = 0$. Moreover, since $F(\cdot)$ is strictly increasing, it follows that $\theta > 0$ (respectively, < 0) and $\mathbb{E}\left[Y_{i2} - Y_{i1} \,|\, X_{i1} = 0\right] > 0$ (resp., < 0) are equivalent. This implies that $\text{sign}(\theta) = \text{sign}\left(\mathbb{E}\left[Y_{i2} - Y_{i1} \,|\, X_{i1} = 0\right]\right)$. A similar argument applied to $X_{i1} = 1$ implies that $\text{sign}(\theta) = \text{sign}\left(\mathbb{E}\left[Y_{i1} - Y_{i2} \,|\, X_{i1} = 1\right]\right)$.

2.8.4 Identification in the exponential model

Let

$$\overline{\phi}_{x_1}(\widetilde{\theta}) \overset{def}{\equiv} \mathbb{E}[\phi_{x_1}(Y_1, Y_2, X_2; \widetilde{\theta}) \,|\, X_{i1} = x_1] = \mathbb{E}[(1 - Y_{i2})e^{\widetilde{\theta}X_{i2}} - (1 - Y_{i1})e^{\widetilde{\theta}X_{i1}} \,|\, X_{i1} = x_1].$$

We show that θ is the unique solution to the equation

$$\overline{\phi}_{x_1}(\widetilde{\theta}) = 0.$$

Since $\overline{\phi}_{x_1}(\theta) = 0$, the result will follow if one can show that, for any $x_1 \in \{0, 1\}$, $\overline{\phi}_{x_1}$ is strictly monotonic.

Let $(\theta_1, \theta_2) \in \Theta^2$ with $\theta_1 > \theta_2$. For $x_1 = 0$, we have

$$\begin{aligned}
&\overline{\phi}_0(\theta_1) - \overline{\phi}_0(\theta_2) \\
&= \mathbb{E}[(1 - Y_{i2})e^{\theta_1 X_{i2}} - (1 - Y_{i1}) \,|\, X_{i1} = 0] - \mathbb{E}[(1 - Y_{i2})e^{\theta_2 X_{i2}} - (1 - Y_{i1}) \,|\, X_{i1} = 0] \\
&= \mathbb{E}[(1 - Y_{i2})(e^{\theta_1 X_{i2}} - e^{\theta_2 X_{i2}}) \,|\, X_{i1} = 0] \\
&= (e^{\theta_1} - e^{\theta_2})\mathbb{E}[(1 - Y_{i2})X_{i2} \,|\, X_{i1} = 0] \\
&= (e^{\theta_1} - e^{\theta_2})\mathbb{E}[(1 - F(\theta + \alpha_i))X_{i2} \,|\, X_{i1} = 0] \\
&= \underbrace{(e^{\theta_1} - e^{\theta_2})}_{>0} \int_{\mathcal{S}} \sum_{y_1=0}^{1} \underbrace{(1 - F(\theta + \alpha))G^2_{y_1,0}(\alpha)F(\alpha)^{y_1}(1 - F(\alpha))^{1-y_1}\pi_0(\alpha)}_{>0 \text{ by Assumption 5}} d\mu(\alpha)
\end{aligned}$$

$$> 0,$$

which shows that $\overline{\phi}_0$ is strictly increasing. If $x_1 = 1$, then

$$\begin{aligned}
&\overline{\phi}_1(\theta_1) - \overline{\phi}_1(\theta_2) \\
&= \mathbb{E}[(1 - Y_{i2})e^{\theta_1 X_{i2}} - (1 - Y_{i1})e^{\theta_1} \,|\, X_{i1} = 1] - \mathbb{E}[(1 - Y_{i2})e^{\theta_2 X_{i2}} - (1 - Y_{i1})e^{\theta_2} \,|\, X_{i1} = 1] \\
&= \mathbb{E}[(1 - Y_{i2})(e^{\theta_1 X_{i2}} - e^{\theta_2 X_{i2}}) - (1 - Y_{i1})(e^{\theta_1} - e^{\theta_2}) \,|\, X_{i1} = 1] \\
&= (e^{\theta_1} - e^{\theta_2})\mathbb{E}[(1 - Y_{i2})X_{i2} - (1 - Y_{i1}) \,|\, X_{i1} = 1] \\
&= -(e^{\theta_1} - e^{\theta_2})\mathbb{E}[(1 - F(\theta + \alpha_i))(1 - X_{i2}) \,|\, X_{i1} = 1] \\
&= -\underbrace{(e^{\theta_1} - e^{\theta_2})}_{>0} \times \\
&\int_{\mathcal{S}} \sum_{y_1=0}^{1} \underbrace{(1 - F(\theta + \alpha))(1 - G^2_{y_1,1}(\alpha))F(\theta + \alpha)^{y_1}(1 - F(\theta + \alpha))^{1-y_1}\pi_1(\alpha)}_{>0 \text{ by Assumption 5}} d\mu(\alpha) \\
&< 0,
\end{aligned}$$

which shows that $\overline{\phi}_1$ is strictly decreasing.

2.8.5 Proof of Lemma 13

In what follows we assume $T \geq 3$, having already proved the validity of the claim for $T = 2$ in Lemma 12.

Since θ is point-identified it is locally point-identified. Additionally, since (θ, π, G) is a regular point of $\nabla Q(\theta, \pi, G)$ by Assumption 6, we can appeal to Theorem 8 in Bekker and Wansbeek (2001) and follow the same line of arguments as in the proof of Lemma 12 to conclude that there exists $x_1 \in \{0, 1\}$ and a $2^{2T-1} \times 1$ vector $\phi_{x_1} \neq 0$ such that

$$\text{(i)} \quad \nabla_{\pi_{x_1}} Q'_{x_1} \phi_{x_1} = 0, \quad \text{(ii)} \quad \nabla_{G_{x_1}} Q'_{x_1} \phi_{x_1} = 0.$$

We will now prove (2.10) and (2.11) using finite induction.
Let us start with (2.10). Given $s \in \{0, ..., T-2\}$, let $\mathcal{P}(s)$ denote the statement that, for all $y^{T-(s+1)} \in \{0, 1\}^{T-(s+1)}$ and $x^{T-(s+1)} \in \{0, 1\}^{T-(s+1)}$,

$$\sum_{y^{T-s:T} \in \{0,1\}^{s+1}} \phi_{x_1}(y^T, x^{2:T}) \prod_{t=T-s}^{T} F(\theta x_t + \alpha)^{y_t}[1 - F(\theta x_t + \alpha)]^{1-y_t}$$

does not depend on $x^{T-s:T}$.

Base case:

Condition (ii) implies that

$$\left(\frac{\partial Q_{x_1}}{\partial G^T_{y^{T-1},x^{T-1}}(\alpha)}\right)' \phi_{x_1} = 0,$$

or equivalently that

$$\begin{aligned}
&\sum_{y_T=0}^{1}\sum_{x_T=0}^{1} \phi_{x_1}(y^T, x^{2:T}) F(\theta x_T + \alpha)^{y_T}[1 - F(\theta x_T + \alpha)]^{1-y_T}(-1)^{1-x_T} \\
&\times \prod_{t=2}^{T-1} F(\theta x_t + \alpha)^{y_t}[1 - F(\theta x_t + \alpha)]^{1-y_t} G^t_{y^{t-1},x^{t-1}}(\alpha)^{x_t}[1 - G^t_{y^{t-1},x^{t-1}}(\alpha)]^{1-x_t} \\
&\times F(\theta x_1 + \alpha)^{y_1}[1 - F(\theta x_1 + \alpha)]^{1-y_1} = 0.
\end{aligned}$$

Using Assumption 5, this simplifies to

$$\sum_{y_T=0}^{1}\sum_{x_T=0}^{1} \phi_{x_1}(y^T, x^{2:T}) F(\theta x_T + \alpha)^{y_T}[1 - F(\theta x_T + \alpha)]^{1-y_T}(-1)^{1-x_T} = 0,$$

which implies that

$$\sum_{y_T=0}^{1} \phi_{x_1}(y^T, x^{2:T}) F(\theta x_T + \alpha)^{y_T}[1 - F(\theta x_T + \alpha)]^{1-y_T}$$

does not depend on x_T.
Thus, $\mathcal{P}(0)$ is true.

Induction step:
Suppose that $\mathcal{P}(0), \ldots, \mathcal{P}(s)$ are true for $s \in \{0, \ldots, T-3\}$. We are going to show that $\mathcal{P}(s+1)$ is true.
Condition (ii) implies that

$$\left(\frac{\partial Q_{x_1}}{\partial G^{T-(s+1)}_{y^{T-(s+2)},x^{T-(s+2)}}(\alpha)}\right)' \phi_{x_1} = 0.$$

If $s < (T-3)$, this corresponds to

$$\begin{aligned}
&\sum_{y^{T-(s+1):T} \in \{0,1\}^{s+2}} \sum_{x^{T-(s+1):T} \in \{0,1\}^{s+2}} \phi_{x_1}(y^T, x^{2:T}) \\
&\times \prod_{t=T-s}^{T} F(\theta x_t + \alpha)^{y_t}[1 - F(\theta x_t + \alpha)]^{1-y_t} G^t_{y^t,x^t}(\alpha)^{x_t}[1 - G^t_{y^t,x^t}(\alpha)]^{1-x_t}
\end{aligned}$$

$$\times F(\theta x_{T-(s+1)}+\alpha)^{y_{T-(s+1)}}[1-F(\theta x_{T-(s+1)}+\alpha)]^{1-y_{T-(s+1)}}(-1)^{1-x_{T-(s+1)}}$$
$$\times \prod_{t=2}^{T-(s+2)} F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t} G^t_{y^{t-1},x^{t-1}}(\alpha)^{x_t}[1-G^t_{y^{t-1},x^{t-1}}(\alpha)]^{1-x_t}$$
$$\times F(\theta x_1+\alpha)^{y_1}[1-F(\theta x_1+\alpha)]^{1-y_1}=0.$$

While if $s=(T-3)$, this corresponds to

$$\sum_{y^{2:T}\in\{0,1\}^{T-1}}\sum_{x^{2:T}\in\{0,1\}^{T-1}}\phi_{x_1}(y^T,x^{2:T})$$
$$\times\prod_{t=3}^{T}F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}G^t_{y^t,x^t}(\alpha)^{x_t}[1-G^t_{y^t,x^t}(\alpha)]^{1-x_t}$$
$$\times F(\theta x_2+\alpha)^{y_2}[1-F(\theta x_2+\alpha)]^{1-y_2}(-1)^{1-x_2}$$
$$\times F(\theta x_1+\alpha)^{y_1}[1-F(\theta x_1+\alpha)]^{1-y_1}=0.$$

Using Assumption 5 this gives, for all $s\in\{0,\dots,T-3\}$,

$$\sum_{y^{T-(s+1):T}\in\{0,1\}^{s+2}}\sum_{x^{T-(s+1):T}\in\{0,1\}^{s+2}}\phi_{x_1}(y^T,x^{2:T})$$
$$\times\prod_{t=T-s}^{T}F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}G^t_{y^t,x^t}(\alpha)^{x_t}[1-G^t_{y^t,x^t}(\alpha)]^{1-x_t}$$
$$\times F(\theta x_{T-(s+1)}+\alpha)^{y_{T-(s+1)}}[1-F(\theta x_{T-(s+1)}+\alpha)]^{1-y_{T-(s+1)}}(-1)^{1-x_{T-(s+1)}}=0. \tag{2.29}$$

Let L_{s+1} denote the left-hand side of (2.29). Exploiting successively the fact that $\mathcal{P}(0),\dots,\mathcal{P}(s)$ are true, alongside the property that, for all $t\in\{T-s,...,T\}$,

$$\sum_{x_t=0}^{1}G^t_{y^t,x^t}(\alpha)^{x_t}[1-G^t_{y^t,x^t}(\alpha)]^{1-x_t}=1, \tag{2.30}$$

it is easy to see that

$$L_{s+1}=\sum_{y^{T-(s+1):T}\in\{0,1\}^{s+2}}\sum_{x_{T-(s+1)}=0}^{1}\phi_{x_1}(y^T,x^{2:T})\prod_{t=T-s}^{T}F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}$$
$$\times F(\theta x_{T-(s+1)}+\alpha)^{y_{T-(s+1)}}[1-F(\theta x_{T-(s+1)}+\alpha)]^{1-y_{T-(s+1)}}(-1)^{1-x_{T-(s+1)}}=0.$$

Recalling that $\mathcal{P}(s)$ is true, this implies that

$$\sum_{y^{T-(s+1):T}\in\{0,1\}^{s+1}}\phi_{x_1}(y^T,x^{2:T})\prod_{t=T-(s+1)}^{T}F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}$$

does not depend on $x^{T-(s+1):T}$. Hence, $\mathcal{P}(s+1)$ is true. This concludes the proof of (2.10).

Finally, we show (2.11). As in the proof of Lemma 12, Condition (i) implies that

$$\begin{aligned}&\sum_{y^T\in\{0,1\}^T}\sum_{x^{2:T}\in\{0,1\}^{T-1}}\phi_{x_1}(y^T,x^{2:T})\\&\times\prod_{t=2}^{T}F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}G^t_{y^t,x^t}(\alpha)^{x_t}[1-G^t_{y^t,x^t}(\alpha)]^{1-x_t}\\&\times F(\theta x_1+\alpha)^{y_1}[1-F(\theta x_1+\alpha)]^{1-y_1}=0.\end{aligned}$$

Using (2.10) and (2.30), it follows that

$$\sum_{y^T\in\{0,1\}^T}\phi_{x_1}(y^T,x^{2:T})\prod_{t=1}^{T}F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}=0,$$

which coincides with (2.11).

2.8.6 Proof of Corollary 13.1

In what follows we assume $T\geq 3$, having already proved the validity of the claim for $T=2$ in Corollary 12.1.

The proof is by contradiction. Suppose that θ is point-identified. We will show that this necessarily leads to $\phi_{x_1}=0$, which will contradict Lemma 13. To that end, we will first prove via finite induction that ϕ_{x_1} must be a constant function.

For $s\in\{1,...,T-2\}$, let $\mathcal{P}(s)$ denote the statement that there exists a function $\phi^{T-s}_{x_1}:\{0,1\}^{2T-2s-1}\to\mathbb{R}$ such that, for all $y^T\in\{0,1\}^T$ and $x^{2:T}\in\{0,1\}^{T-1}$, we have

$$\phi_{x_1}(y^T,x^{2:T})=\phi^{T-s}_{x_1}(y^{T-s},x^{2:T-s}).$$

Base case:
By (2.10), the quantity

$$\sum_{y_T=0}^{1}\phi_{x_1}(y^T,x^{2:T})F(\theta x_T+\alpha)^{y_T}[1-F(\theta x_T+\alpha)]^{1-y_T}\tag{2.31}$$

does not depend on x_T. Hence

$$\phi_{x_1}(y^{T-1}, 1, x^{2:T-1}, 1)F(\theta+\alpha) + \phi_{x_1}(y^{T-1}, 0, x^{2:T-1}, 1)[1-F(\theta+\alpha)]$$
$$= \phi_{x_1}(y^{T-1}, 1, x^{2:T-1}, 0)F(\alpha) + \phi_{x_1}(y^{T-1}, 0, x^{2:T-1}, 0)[1-F(\alpha)].$$

By linear independence of 1, $F(\alpha)$, and $F(\theta+\alpha)$, this implies that $\phi_{x_1}(y^T, x^{2:T})$ does not depend on (y_T, x_T). Hence $\mathcal{P}(1)$ is true.

Induction step

Suppose that $\mathcal{P}(s)$ is true for $s \in \{1, ..., T-3\}$. Let us show that $\mathcal{P}(s+1)$ is true.

Since $\mathcal{P}(s)$ is true, we know that there exists a function $\phi_{x_1}^{T-s} : \{0,1\}^{2T-2s-1} \to \mathbb{R}$ such that

$$\phi_{x_1}(y^T, x^{2:T}) = \phi_{x_1}^{T-s}(y^{T-s}, x^{2:T-s}).$$

Thus, by (2.10), the quantity:

$$\sum_{y^{T-s:T} \in \{0,1\}^{s+1}} \phi_{x_1}(y^T, x^{2:T}) \prod_{t=T-s}^{T} F(\theta x_t + \alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}$$
$$= \sum_{y_{T-s}=0}^{1} \phi_{x_1}^{T-s}(y^{T-s}, x^{2:T-s}) \sum_{y^{T-(s-1):T} \in \{0,1\}^s} \prod_{t=T-(s-1)}^{T} F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}$$
$$\times F(\theta x_{T-s}+\alpha)^{y_{T-s}}[1-F(\theta x_{T-s}+\alpha)]^{1-y_{T-s}}$$
$$= \sum_{y_{T-s}=0}^{1} \phi_{x_1}^{T-s}(y^{T-s}, x^{2:T-s})F(\theta x_{T-s}+\alpha)^{y_{T-s}}[1-F(\theta x_{T-s}+\alpha)]^{1-y_{T-s}}$$

does not depend on $x^{T-s:T}$. Therefore,

$$\phi_{x_1}^{T-s}(y^{T-s-1}, 1, x^{2:T-s-1}, 1)F(\theta+\alpha) + \phi_{x_1}^{T-s}(y^{T-s-1}, 0, x^{2:T-s-1}, 1)[1-F(\theta+\alpha)]$$
$$= \phi_{x_1}^{T-s}(y^{T-s-1}, 1, x^{2:T-s-1}, 0)F(\alpha) + \phi_{x_1}^{T-s}(y^{T-s-1}, 0, x^{2:T-s-1}, 0)[1-F(\alpha)].$$

Since 1, $F(\alpha)$, and $F(\theta+\alpha)$ are linearly independent, this implies $\mathcal{P}(s+1)$.
It follows from the previous induction argument that there exists a function $\phi_{x_1}^2 : \{0,1\}^3 \to \mathbb{R}$ such that, for all $(y^T, x^{2:T})$,

$$\phi_{x_1}(y^T, x^{2:T}) = \phi_{x_1}^2(y^2, x_2).$$

Using (2.10), the quantity

$$\sum_{y^{2:T} \in \{0,1\}^{T-1}} \phi_{x_1}(y^T, x^{2:T}) \prod_{t=2}^{T} F(\theta x_t+\alpha)^{y_t}[1-F(\theta x_t+\alpha)]^{1-y_t}$$

$$= \sum_{y_2=0}^{1} \phi_{x_1}^2(y^2, x_2) F(\theta x_2 + \alpha)^{y_2} [1 - F(\theta x_2 + \alpha)]^{1-y_2}$$

does not depend on $x^{2:T}$. Therefore,

$$\begin{aligned} &\phi_{x_1}^2(y_1, 1, 1) F(\theta + \alpha) + \phi_{x_1}^2(y_1, 0, 1)[1 - F(\theta + \alpha)] \\ &= \phi_{x_1}^2(y_1, 1, 0) F(\alpha) + \phi_{x_1}^2(y_1, 0, 0)[1 - F(\alpha)]. \end{aligned}$$

Since 1, $F(\alpha)$, and $F(\theta+\alpha)$ are linearly independent, this implies that there exists a function $\phi_{x_1}^1 : \{0, 1\} \to \mathbb{R}$ such that, for all $(y^T, x^{2:T})$,

$$\phi_{x_1}(y^T, x^{2:T}) = \phi_{x_1}^1(y_1).$$

Lastly, (2.11) implies

$$\begin{aligned} &\sum_{y^T \in \{0,1\}^T} \phi_{x_1}(y^T, x^{2:T}) \prod_{t=1}^{T} F(\theta x_t + \alpha)^{y_t} [1 - F(\theta x_t + \alpha)]^{1-y_t} \\ &= \sum_{y^T \in \{0,1\}^T} \phi_{x_1}^1(y_1) \prod_{t=1}^{T} F(\theta x_t + \alpha)^{y_t} [1 - F(\theta x_t + \alpha)]^{1-y_t} \\ &= \sum_{y_1=0}^{1} \phi_{x_1}^1(y_1) \sum_{y^{2:T} \in \{0,1\}^T} \prod_{t=1}^{T} F(\theta x_t + \alpha)^{y_t} [1 - F(\theta x_t + \alpha)]^{1-y_t} \\ &= \sum_{y_1=0}^{1} \phi_{x_1}^1(y_1) F(\theta x_1 + \alpha)^{y_1} [1 - F(\theta x_1 + \alpha)]^{1-y_1} \\ &= 0. \end{aligned}$$

Linear independence of 1, $F(\alpha)$, and $F(\theta + \alpha)$ thus implies

$$\phi_{x_1}^1(0) = \phi_{x_1}^1(1) = 0.$$

Therefore, ϕ_{x_1} must be the null function, a contradiction.

2.8.7 Proof of Proposition 5

It is immediate to verify that, if $\widetilde{\theta} \in \Theta^I$, then (2.16), (2.17) and (2.18) are satisfied. Conversely, suppose that (2.16), (2.17) and (2.18) are satisfied. Let

$$p_{x_1}(y^T, x^{2:T}, \alpha) = F(\widetilde{\theta} x_T + \alpha)^{y_T} [1 - F(\widetilde{\theta} x_T + \alpha)]^{1-y_T} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha). \tag{2.32}$$

Using (2.17) we have

$$p_{x_1}(y^T, x^{2:T}, \alpha) \geq 0, \qquad \sum_{y^T \in \{0,1\}^T} \sum_{x^{2:T} \in \{0,1\}^{T-1}} \int_{\mathcal{S}} p_{x_1}(y^T, x^{2:T}, \alpha) d\mu(\alpha) = 1,$$

so p_{x_1} is a valid distribution function (conditional on $X_{i1} = x_1$).
Next, using (2.16) we have

$$\begin{aligned}
&\int_{\mathcal{S}} p_{x_1}(y^T, x^{2:T}, \alpha) d\mu(\alpha) \\
&= \int_{\mathcal{S}} F(\widetilde{\theta} x_T + \alpha)^{y_T} [1 - F(\widetilde{\theta} x_T + \alpha)]^{1-y_T} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) d\mu(\alpha) \\
&= Q_{x_1}(y^T, x^{2:T}; \theta, \pi, G),
\end{aligned}$$

so p_{x_1} is consistent with the conditional distribution $Q_{x_1}(y^T, x^{2:T}; \theta, \pi, G)$ of $(Y_i^T, X_i^{2:T})$ given X_{i1}.
Next, using (2.18) we have, for all $s \in \{2, ..., T\}$,

$$\begin{aligned}
&\sum_{x^{s:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s:T} \in \{0,1\}^{T-s+1}} p_{x_1}(y^T, x^{2:T}, \alpha) \\
&= \sum_{x^{s:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s:T-1} \in \{0,1\}^{T-s}} \left\{ \sum_{y_T=0}^{1} F(\widetilde{\theta} x_T + \alpha)^{y_T} [1 - F(\widetilde{\theta} x_T + \alpha)]^{1-y_T} \right\} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) \\
&= \sum_{x^{s:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s:T-1} \in \{0,1\}^{T-s}} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) \\
&= F(\widetilde{\theta} x_{s-1} + \alpha)^{y_{s-1}} [1 - F(\widetilde{\theta} x_{s-1} + \alpha)]^{1-y_{s-1}} \sum_{x^{s:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s-1:T-1} \in \{0,1\}^{T-s+1}} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) \\
&= F(\widetilde{\theta} x_{s-1} + \alpha)^{y_{s-1}} [1 - F(\widetilde{\theta} x_{s-1} + \alpha)]^{1-y_{s-1}} \sum_{x^{s:T} \in \{0,1\}^{T-s+1}} \sum_{y^{s-1:T} \in \{0,1\}^{T-s+2}} p_{x_1}(x^{2:T}, y^T, \alpha),
\end{aligned}$$

so, for all $t \in \{1, ..., T-1\}$, the conditional distributions of Y_{it} given $(Y_i^{t-1}, X_i^{t-1}, \alpha_i)$ induced by p_{x_1} coincide with the ones under the model; i.e., with $F(\widetilde{\theta} x_t + \alpha)^{y_t} [1 - F(\widetilde{\theta} x_t + \alpha)]^{1-y_t}$.
Lastly, using (2.32) we have

$$\begin{aligned}
p_{x_1}(y^T, x^{2:T}, \alpha) &= F(\widetilde{\theta} x_T + \alpha)^{y_T} [1 - F(\widetilde{\theta} x_T + \alpha)]^{1-y_T} \psi_{x_1}(x^{2:T}, y^{T-1}, \alpha) \\
&= F(\widetilde{\theta} x_T + \alpha)^{y_T} [1 - F(\widetilde{\theta} x_T + \alpha)]^{1-y_T} \sum_{y_T=0}^{1} p_{x_1}(y^T, x^{2:T}, \alpha),
\end{aligned}$$

so the conditional distribution of Y_{iT} given $(Y_i^{T-1}, X_i^{T-1}, \alpha_i)$ induced by p_{x_1} also coincides with the one under the model.
This implies that $\widetilde{\theta} \in \Theta^I$.

2.8.8 Computation of identified sets

In this section we describe the practical implementation of the linear programming approach for the computation of identified sets for two types of target parameters: θ, and average

partial effects. For simplicity of exposition we discuss the case $T = 2$, but the construction is analogous for larger T.

2.8.8.1 Parameter θ

In Proposition 5, we established that a candidate parameter $\widetilde{\theta}$ lies in the identified set Θ^I if and only if one can find functions ψ_0, ψ_1 verifying equations (2.13), (2.14) and (2.15). A useful observation is that these conditions can be viewed as the constraints of a linear program. Thus, determining whether $\widetilde{\theta} \in \Theta^I$ is equivalent to determining the feasibility of a linear optimization problem. In the numerical illustration, we specifically consider:

$$\inf_{\psi_0,\psi_1} \int_{\mathcal{S}} \sum_{x_1=0}^{1} q_{x_1} \sum_{x_2=0}^{1} \sum_{y_1=0}^{1} \psi_{x_1}(x_2, y_1, \alpha) d\mu(\alpha),$$

where the constraints are that ψ_0, ψ_1 satisfy equations (2.13), (2.14) and (2.15). The additional constraints for the strictly exogenous case are that ψ_0, ψ_1 also verify the relationship presented in footnote 5.

2.8.8.2 Average partial effect Δ

In addition to θ, a quantity of interest is the average partial effect

$$\begin{aligned}\Delta &= \mathbb{E}[\Pr\big(Y_{i2} = 1 \,|\, X_{i2} = 1, \alpha_i\big) - \Pr\big(Y_{i2} = 1 \,|\, X_{i2} = 0, \alpha_i\big)] \\ &= \int_{\mathcal{S}} [F(\theta + \alpha) - F(\alpha)] \sum_{x_1 \in \{0,1\}} q_{x_1} \pi_{x_1}(\alpha) d\mu(\alpha).\end{aligned}$$

which is generally not point-identified. Yet, for a given $\widetilde{\theta} \in \Theta^I$, one can compute a lower bound $\underline{\Delta}(\widetilde{\theta})$ and an upper bound $\overline{\Delta}(\widetilde{\theta})$ on the range of possible average partial effects as solutions to the following linear optimization problem:

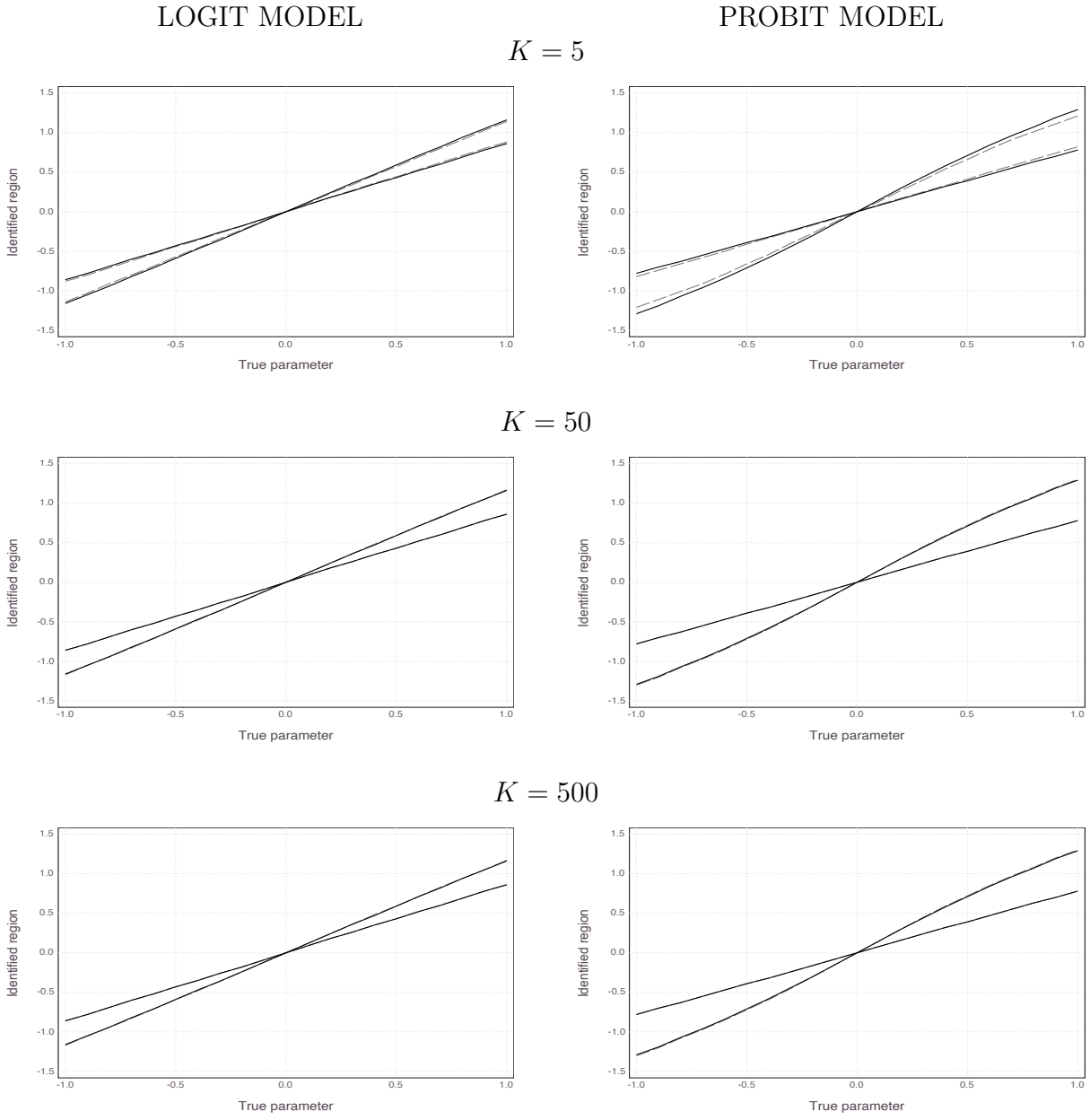
$$\underline{\Delta}(\widetilde{\theta}) = \inf_{\psi_0,\psi_1} \int_{\mathcal{S}} [F(\widetilde{\theta} + \alpha) - F(\alpha)] \sum_{x_1 \in \{0,1\}} q_{x_1} \sum_{x_2 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \psi_{x_1}(x_2, y_1, \alpha) d\mu(\alpha),$$

$$\overline{\Delta}(\widetilde{\theta}) = \sup_{\psi_0,\psi_1} \int_{\mathcal{S}} [F(\widetilde{\theta} + \alpha) - F(\alpha)] \sum_{x_1 \in \{0,1\}} q_{x_1} \sum_{x_2 \in \{0,1\}} \sum_{y_1 \in \{0,1\}} \psi_{x_1}(x_2, y_1, \alpha) d\mu(\alpha),$$

subject to ψ_0, ψ_1 satisfying equations (2.15), (2.13), and (2.14). Under the assumption of strict exogeneity, ψ_0 and ψ_1 have to satisfy the additional constraint discussed in footnote 5. The sharp bounds for Δ are then obtained as

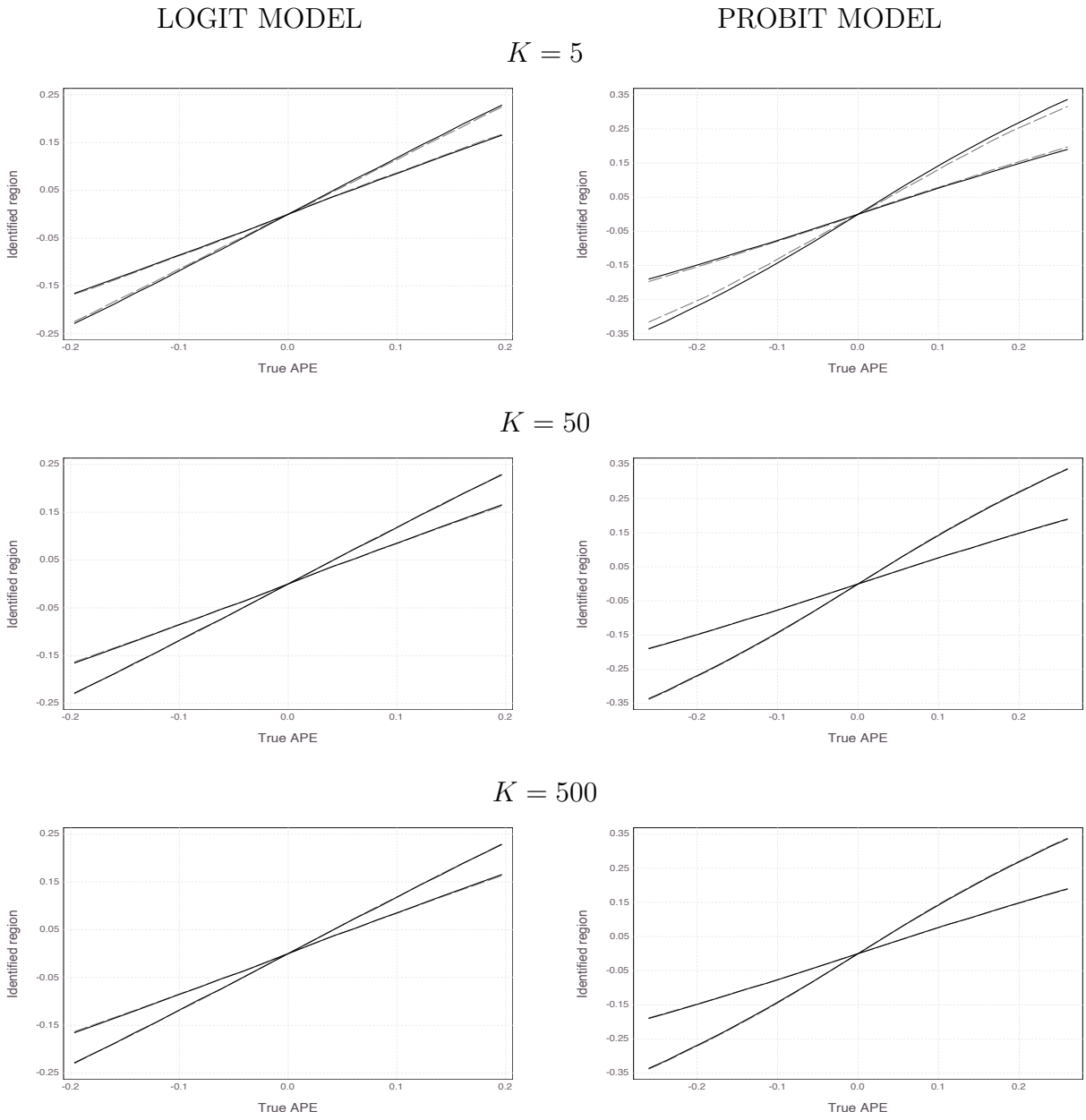
$$\begin{aligned}\underline{\Delta} &= \inf_{\widetilde{\theta} \in \Theta^I} \underline{\Delta}(\widetilde{\theta}), \\ \overline{\Delta} &= \sup_{\widetilde{\theta} \in \Theta^I} \overline{\Delta}(\widetilde{\theta}).\end{aligned}$$

Figure 2.3: Approximate identified sets for logit and probit models with $T = 2$



Notes: Approximate upper and lower bounds of the identified set Θ^I in a logit model (left column) and a probit model (right column) with $T = 2$ based on a discretization of unobserved heterogeneity with $K = 5, 50, 500$ support points respectively. The true identified set is depicted by the solid lines while the approximations are indicated by the dashed lines. The population value of θ is given on the x-axis.

Figure 2.4: Approximate identified sets for average partial effects in logit and probit models with $T = 2$



Notes: Approximate upper and lower bounds of the identified set for average partial effects in a logit model (left column) and a probit model (right column) with $T = 2$ using a discretization of unobserved heterogeneity with $K = 5, 50, 500$ support points respectively. The true identified set is depicted by the solid lines while the approximations are indicated by the dashed lines. The population value is given on the x-axis.

Chapter 3

Identification and estimation of random effects linear social interaction models with endogenous peer selection

3.1 Introduction

Most studies on peer effects treat friendships between individuals either as fixed or as exogenous random variables in a linear social interaction model for individual outcomes. There are reasons to be doubtful of this convention; namely the possibility that unobserved individual traits such as personality type or sociability level may influence both peer selection (e.g Selfhout et al. (2010)) and individual outcomes (e.g Golsteyn et al. (2021) in the context of academic achievement). A concrete example is helpful to flesh out the econometric problem. Suppose as in Manski (1993) that a researcher is interested in measuring the effects of an educational intervention providing tutoring program to some students and not to others on, say, student GPA. Let $i = 1, ..., N$ index students in the school and assume that the researcher observes an undirected sociomatrix D, where the ij^{th} entry $D_{ij} \in \{0, 1\}$ denotes a potential friendship between students i and j. The binary treatment X_i is randomly assigned amongst students and a simple *linear-in-means* specification is considered to account for potential spillover effects of the treatment on the outcome of interest Y_i:

$$Y_i = \alpha_0 + \gamma_0 X_i + \delta_0 \sum_{j \neq i} G_{ij} X_j + U_i \tag{3.1}$$

Here, $G = (G_{ij})_{i,j=1}^{N}$ is the row-normalized version of D:

$$G_{ij} = \begin{cases} \frac{D_{ij}}{D_{i+}} & \text{if } D_{i+} > 0 \\ 0 & \text{otherwise} \end{cases}$$

with $D_{i+} = \sum_{j \neq i} D_{ij}$ denoting student i's degree centrality, i.e his number of friends. The error term U_i can be regarded as a latent sociability level of student i. By virtue of random assignment, X_i is an exogenous covariate but the fraction of friends treated, $\sum_{j \neq i} G_{ij} X_j$, is not due to the potential correlation between the latent sociability of individual i, U_i, and the identity of her friends. To see this, observe the following:

$$
\begin{aligned}
& Cov(U_i, \sum_{j \neq i} G_{ij} X_j) \\
&= \sum_{j \neq i} Cov(U_i, G_{ij} X_j) \\
&= \sum_{j \neq i} \mathbb{E}(U_i G_{ij} X_j) - \mathbb{E}(U_i)\mathbb{E}(G_{ij} X_j) \text{ (by linearity)} \\
&= \sum_{j \neq i} \left(\mathbb{E}(U_i G_{ij}) - \mathbb{E}(U_i)\mathbb{E}(G_{ij}) \right) \mathbb{E}(X_j) \text{ (by random assignment)} \\
&= \mathbb{E}(X_1) \left(\mathbb{E}(U_i \sum_{j \neq i} G_{ij}) - \mathbb{E}(U_i)\mathbb{E}(\sum_{j \neq i} G_{ij}) \right) \\
&= \mathbb{E}(X_1) \left(\mathbb{E}\left(U_i \sum_{j \neq i} \frac{D_{ij}}{D_{i+}} \mathbb{1}\{D_{i+} > 0\} \right) - \mathbb{E}(U_i)\mathbb{E}\left(\sum_{j \neq i} \frac{D_{ij}}{D_{i+}} \mathbb{1}\{D_{i+} > 0\} \right) \right) \\
&= \mathbb{E}(X_1) \left(\mathbb{E}\left(U_i \underbrace{\sum_{j \neq i} \frac{D_{ij}}{D_{i+}}}_{=1} \middle| D_{i+} > 0 \right) - \mathbb{E}(U_i)\mathbb{E}\left(\underbrace{\sum_{j \neq i} \frac{D_{ij}}{D_{i+}}}_{=1} \middle| D_{i+} > 0 \right) \right) P(D_{i+} > 0) \\
&= \mathbb{E}(X_1) \left(\mathbb{E}(U_i | D_{i+} > 0) - \mathbb{E}(U_i) \right) P(D_{i+} > 0)
\end{aligned}
$$

Generally, we will have $\mathbb{E}(U_i | D_{i+} > 0) \neq \mathbb{E}(U_i)$ as we expect positive degree-centrality to be indicative of sociability level in networks of finite size. In turn, a standard regression will lead to biased estimates of the spillover effect δ_0, the size of which will depend on the precise features of the model.

This chapter presents an approach to deal with this issue. Our work fits in the growing literature initiated by Goldsmith-Pinkham and Imbens (2013) that aims to formally account for network endogeneity in peer effect models by jointly modelling link formation and individual outcomes. Motivated by the above example, we assume that individuals have latent attributes that influence outcomes and exclusively determine friendships according to a conditionally independent dyad model in the spirit of Auerbach (2019), Johnsson and Moon (2015) and Shalizi and McFowland III (2016). However, our methodology and assumptions are substantially different from these papers which focus on estimation in "large" networks while we focus on the more empirically motivated "small" network setting wherein

the researcher observes several independent clusters of moderate size, e.g. classrooms, teams, schools. In that respect, our framework is closer to Jochmans (2020) and we develop complementary approaches, though to our knowledge, we are the first to address the additional complication of correlated unobservables and latent exogenous effects in this setup. Otherwise, whereas Jochmans (2020) analyzes directed networks, and uses stochastic restrictions of the link formation process to build instrumental variables, we study undirected networks, and exploit inherent symmetries of the model to formulate a different solution to the problem of network endogeneity. Specifically, leveraging the exchangeability of the link formation model and an independence assumption between observable and latent characteristics, we show that controlling or matching individuals by degree-centrality is sufficient to eliminate the omitted variable bias induced by endogenous peer selection. We combine this result and insights from Bramoullé et al. (2009) for the case of exogenous friendships to propose two simple strategies for the identification and estimation of social effects. Finally, we draw on the recent work of Hansen and Lee (2019), to show that our estimators are consistent and asymptotically normal under standard assumptions for clustered samples.

The remainder of the chapter is organized as follows. Section 3.2 introduces the model and our working assumptions. Section 3.3 further discusses the complications arising from endogenous peer selection. In Section 3.4, 3.5 and 3.6 we introduce our estimators and characterize their properties. We present Monte Carlo results in Section 3.7 and Section 3.8 concludes. Proofs are gathered in the Appendix.

3.2 The econometric model

Our focus is on the estimation of linear social interactions models subject to endogenous peer selection when the researcher observes data from $c = 1, \dots, C$ distinct clusters (e.g schools, classrooms) of finite size. To that end, we hypothesize a common data generating process for each cluster that we detail below.

3.2.1 Data generating process

A cluster c comprises a finite population of N_c agents labelled by the integers in $\{1, \dots, N_c\}$. Each agent i in the cluster is endowed with a pair, $(X_i, U_i) \in \mathbb{R}^K \times \mathbb{R}$ known to her. The K-dimensional vector X_i captures observable individual specific characteristics while U_i is a scalar agent-level attribute unobserved by the econometrician. We collect these features in the vector $U_c = (U_1, \dots, U_{N_c})$ and the matrix $X_c = (X_1, \dots, X_{N_c})$ - the cluster subscript c will be omitted when doing so causes no confusion. Aside from the primitives X, U, cluster variables include an undirected network D and a vector of individual outcomes Y generated in two consecutive stages.

First, we posit a peer selection stage producing the symmetric adjacency matrix D. Specifically, we assume that agents form friendships according to the following non-strategic,

non-parametric dyadic network formation model:

$$D_{ij} = \mathbb{1}\left\{h(U_i, U_j) - V_{ij} \geq 0\right\} \mathbb{1}\left\{i \neq j\right\} \tag{3.2}$$

$$V_{ij} \sim^{iid} F(.),\ V_{ij} \text{ independent of } U$$

Here, $h(.,.)$ is a symmetric function in the latent attributes (U_i, U_j) called a graphon and $V_{ij} = V_{ji}$ are idiosynchratic shocks. In this setup, the existence of a dyad ij (i.e $D_{ij} = 1$) depends on whether the total surplus of forming a friendship, $h(U_i, U_j) - V_{ij}$, exceeds the zero threshold. Observe that the link probability: $P(D_{ij} = 1 | U_i = u_i, U_j = u_j) = (F \circ h)(u_i, u_j)$ is an increasing function of (u_i, u_j) when the graphon is increasing in its arguments. In this case, agents with a comparatively "high" u_i will tend to have a comparatively higher number of friends, i.e a higher degree centrality, and thus we will generally interpret U_i as a latent popularity. Note the important restriction that observable characteristics X_i do not enter link decisions.

In a second stage, given (X, U, D), we assume that all agents in the cluster play a linear-quadratic game resulting in a pure strategy Nash equilibrium Y. We consider two possibilities: the *linear-in-means* specification and the *local-aggregate* specification. In the former, the utility from an action profile y takes the form:

$$\forall i \in \{1, \ldots, N\},\ u_i(y; D, U, X) = v_i(D, U, X) y_i - \frac{1}{2} y_i^2 + \beta_0 \bar{y}_{n(i)} y_i \tag{3.3}$$

where $\bar{y}_{n(i)} = \sum_{j \neq i} G_{ij} Y_j$ denotes the average action i's friends in the network D. Here,

$$v_i(D, U, X) = \alpha_0 + X_i' \gamma_0 + \bar{X}_{n(i)}' \delta_0 + A_D + U_i + \lambda_0 \bar{U}_{n(i)}$$

where $\bar{X}_{n(i)} = \sum_{j \neq i} G_{ij} X_j$, $\bar{U}_{n(i)} = \sum_{j \neq i} G_{ij} U_j$ capture the average observable (respectively unobservable) characteristics of i's peers and A_D is a scalar network effect. Solving for the best response behavior leads to:

$$Y_i = \alpha_0 + \beta_0 \bar{Y}_{n(i)} + X_i' \gamma_0 + \bar{X}_{n(i)}' \delta_0 + A_D + U_i + \lambda_0 \bar{U}_{n(i)} \tag{3.4}$$

Equation (3.4) corresponds to what is widely known as the *linear-in-means* model. As is common in the literature, we impose $|\beta_0| < 1$ to ensure the uniqueness of the pure strategy Nash equilibrium of the game.

The *local-aggregate* model is a variant focusing on aggregate quantities rather than averages. In this model the utility function is:

$$\forall i \in \{1, \ldots, N\},\ u_i(y; D, X) = v_i(D, X) y_i - \frac{1}{2} y_i^2 + \beta_0 \sum_{j \neq i} D_{ij} y_i y_j \tag{3.5}$$

$$v_i(D, X) = \alpha_0 + X_i' \gamma_0 + \left(\sum_{j \neq i} D_{ij} X_j \right)' \delta_0 + A_D + U_i + \lambda_0 \sum_{j \neq i} D_{ij} U_j$$

Then, the first order condition for optimality imply a linear best reply of the form:

$$Y_i = \alpha_0 + \beta_0 \sum_{j \neq i} D_{ij} Y_j + X_i' \gamma_0 + \left(\sum_{j \neq i} D_{ij} X_j \right)' \delta_0 + A_D + U_i + \lambda_0 \sum_{j \neq i} D_{ij} U_j \tag{3.6}$$

Similarly, we will assume $(N-1)|\beta_0| < 1$, a sufficient condition to ensure the uniqueness of the pure strategy Nash equilibrium (Ballester et al. (2006)).

Arguably less popular than its counterpart the *local-aggregate* model has nevertheless been the subject of several studies in game theory (Ballester et al. (2006), Calvó-Armengol et al. (2009), Ushchev and Zenou (2020)), in econometrics (Liu and Lee (2010), Liu et al. (2014)) and in the education literature (Calvó-Armengol et al. (2009), Liu et al. (2014)). The two models can serve to emphasize different facets of peer influence. For instance, the *linear-in-means* model has commonly been employed to represent social conformity in a group of individuals. To understand why, observe that the utility function (3.3) is isomorphic to:

$$u_i(y; D, U, X) = v_i'(D, U, X) y_i - \frac{1}{2} \left(y_i^2 + \zeta_0 \left(y_i - \bar{y}_{n(i)} \right)^2 \right)$$

In this reformulation an individual's utility is affected by the deviation of her action from that of her reference group. If $\zeta_0 > 0$, the agent will try to mimic the mean action of her friends to maximize utility. In contrast, the utility function of the *local-aggregate* model (3.5) highlights the complementary of actions between connected individuals. For this reason, the model has traditionally been used to analyze the role of spillovers in education (Calvó-Armengol et al. (2009), Ushchev and Zenou (2020)). For our purposes and from an econometric viewpoint, the critical difference between the two models is that one employs the row-normalized adjacency matrix while the other uses the adjacency matrix.[1] We will see that this difference has important implications for identification and estimation.

For ease of exposition, we will focus on the scalar case ($K = 1$) and to facilitate the joint treatment of the two specifications, we will work with the following notation:

$$Y_i = \alpha_0 + \beta_0 \sum_{j \neq i} \omega_{ij}(D_i) Y_j + \gamma_0 X_i + \delta_0 \sum_{j \neq i} \omega_{ij}(D_i) X_j + A_D + U_i + \lambda_0 \sum_{j \neq i} \omega_{ij}(D_i) U_j \tag{3.7}$$

where it is understood that $\omega_{ij}(D_i) = G_{ij}$ for the *linear-in-means* and $\omega_{ij}(D_i) = D_{ij}$ for the *local-aggregate* model. Note the rich structure of the composite error term: $\epsilon_i = A_D + U_i + \lambda_0 \sum_{j \neq i} \omega_{ij}(D_i) U_j$, which allows for complex forms of dependence within each cluster going beyond the simple correlation induced by network fixed effects [2].

[1] Another more subtle difference of the *linear-in-sums* is that the equilibrium outcome will be proportional to the Katz-Bonacich centrality of the agent in the network (Ballester et al. (2006), Calvó-Armengol et al. (2009)). This feature is absent in the *linear-in-means* model due to the row-normalization of the adjacency matrix.

[2] Suprisingly, a very few number of papers entertain the possibility of latent exogenous effect; Graham (2008) and Graham et al. (2020) being notable exceptions. This common asymmetric treatment of observables and unobservables covariates effectively adds restrictions to the model that are rarely mentioned.

This data generating process produces a tuple $\{Y, D, X, U, A_D\}$ for each cluster from which only $\{Y, X, D\}$ are observed by the econometrician; U and A_D being latent by definition. Our goal in this chapter is to provide identification conditions and discuss methods to estimate the common vector of (observable) social effects: $\theta_0 = (\beta_0, \gamma_0, \delta_0)' \in \mathbb{R}^{2K+1}$ [3]. We follow the terminology of Manski (1993) and refer to β_0 as the endogenous effect and δ_0 as the exogenous effect. We point out that the fact that U enters linearly in equation (3.4) or (3.6) is for expository purposes and is inconsequential for the subsequent discussion [4].

3.2.2 Discussion of the network model

A key characteristic of network formation model (3.2) that we will exploit throughout is that links form independently conditionally on the vector of agent-specific latent attribute U with:

$$D_{ij}|U_i, U_j \sim \text{Bernouilli}\left((F \circ h)(U_i, U_j)\right)$$

for every dyad $\{i, j\}$, $i < j$. Models that feature this property are called Conditionally Independent Dyad Models (see Graham (2020)) - henceforth CID - and are ubiquitous in the network-related literature: the Erdős–Rényi random graph model (Erdős and Rényi (1960)), the β-model (Chatterjee et al. (2011)) and the Stochastic Block Model (Holland et al. (1983)) being notable special cases. Prior work in economics that employ these models are Auerbach (2019) and Johnsson and Moon (2015) in a similar context to ours.

Note that in its present form, our link formation model (3.2) does not accommodate assortative matching on observable characteristics. This turns out to be a facilitating element for identification of the social effects. However, we discuss how the model may incorporate this behavioral dimension if we impose that covariates entering links decisions are distinct from those featuring in the outcome equation in Section 3.5. This generalization requires a distributional exclusion restriction (Powell (1994), p. 2484)) to preserve the identification of the social effects but has the advantage to allows for some degree of correlation between the latent attributes and the observable characteristics. Finally, we point out that the network modelling approach pursued here has two limitations: strategic considerations (e.g reciprocity, transitivity) are absent, though they may be critical in certain contexts, and we rule out the possibility of a feeback effect of outcomes on link formation.

3.2.3 Sampling and main assumptions

As in Jochmans (2020), we follow the conventional sampling view in the peer effect literature and assume that the researcher has a sample of **independent** networks of finite size. In the language of network inference, we cast our identification and estimation strategies in

[3] The parameter α_0 cannot be identified due to the presence of network fixed effects

[4] The crucial assumption common to all conventional peer effects models is that all terms enter in an additive separable way.

the *many networks* asymptotics. As detailed in Section 3.2.1, we assume the econometrician observes a collection of C clusters/networks (we use the terms interchangeably), potentially of different sizes, with observations $\{Y_c, X_c, D_c\}_{c=1}^{C}$ produced by the same data generating process. Importantly, this entails that each adjacency matrix is generated according to a CID process of the form (3.2) and that outcomes are generated with the same social influence vector: $\theta_0 = (\beta_0, \gamma_0, \delta_0)'$ in either the *linear-in-means* model or the *local-aggregate* model. Our asymptotic framework involves $C \to \infty$.

The following assumptions are assumed to hold for each individual cluster:

Assumption 7. *(IID-ness)*
$\{X_i, U_i\}_{i=1}^{N}$ *are independent and identically distributed with finite second moment. The density of* U_i *denoted* $f_U(u) = f(u)$ *is continuous with support* $\mathcal{U} \subset \mathbb{R}$*. The covariates* X_i *may be discrete or continous and are non degenerate random variables.*

Assumption 8. *(Random Effect)*
$\forall i \in \{1, \ldots, N\} : X_i$ *and* U_i *are independent*

Assumption 9. *(Correlated Unobservables)*
The network fixed effect A_D *is arbitrarily correlated with* X*,*U*.*

Assumption 10. *(CID network model)*

1. *The graphon* $h : \mathcal{U} \times \mathcal{U} \mapsto [0, 1]$ *is symmetric, measurable and non-degenerate*
2. $\{V_{ij}\}_{1 \leq i < j \leq N}$ *follow a standard uniform, are independent across dyads* $\{i, j\}$*, and independent of* (A_D, U, X)*.*

Assumption 8 is arguably restrictive but a fixed effect assumption for U_i is likely to be too demanding in our context. Indeed, since we are working with a cross-section of networks as opposed to repeated observations of the same network over time, differencing out the fixed effect as in linear panel data is problematic. Thus, the alternative is to treat the latent attributes as additional parameters to be estimated from the data. However, we know by analogy with the panel data literature on the incidental parameter problem that this would be equally inadequate in common economic applications where clusters have a relatively small size. This is of course a lesser issue for very large networks and when working within a *large network asymptotics* framework (Johnsson and Moon (2015)) since the effective number of observations (i.e links) per individual is increasing with the sample size. Our restrictions on unobservables would be particularly fitting in the context of a randomized experiment where a treatment X would be randomly assigned and thus independent of unobserved heterogeneity U. We will discuss how Assumption 8 can be relaxed to some extent in Section 3.5. Assumptions 7, 9 and 10 are otherwise standard.

A word on notation and conventions. Hereafter, a lower case letter will denote a specific value of the random variable denoted by the corresponding upper-case letter. We use

the generic notation $f_Y(y)$, $f_{Y,X}(y,x)$, $f_{Y|X}(y|x)$ to denote a density, a joint density and a conditional density. We also use the standard notation $X \sim Y$ to indicate that the random variables X and Y are identically distributed. Finally, to conserve on space, we introduce $D_i = (D_{i1}, \ldots, D_{i(i-1)}, D_{i(i+1)}, \ldots, D_{iN})$ and $G_i = (G_{i1}, \ldots, G_{i(i-1)}, G_{i(i+1)}, \ldots, G_{iN})$ as the ith row of the adjacency matrix D and the row-normalized adjacency matrix G representing a network of size N. We use agent(s), individual(s), node(s) and vertex/vertices interchangeably and likewise for link(s) and edge(s).

3.3 The issue of endogenous peer groups

To understand the complications that arise when individuals self select their peers, we start by considering a trimmed down version of the *linear-in-means* with no network effect and that only features exogenous social effects

$$Y_i = \alpha_0 + \gamma_0 X_i + \delta_0 \sum_{j \neq i} G_{ij} X_j + U_i \tag{3.8}$$

We refer to Equation (3.8) as the baseline *linear-in-means* model. We are interested in the identification of $\theta_0 = (\gamma_0, \delta_0)' \in \mathbb{R}^{2K}$. To help with interpretation, reconsider our introductory example wherein a researcher is interested in the effect of an educational intervention providing tutoring program to certain students on student GPA. He employs a random assignment procedure and uses model (3.8) to account for potential spillover effects. In this case, we can view (3.8) as a linear regression model explaining the GPA of student i in terms of his treatment status X_i, and the fraction of his peers participating in the tutoring program, $\sum_{j \neq i} G_{ij} X_j$. While X_i is an exogenous covariate in this context (i.e Assumption 8), a priori, it is unclear if this is also the case for the fraction of her friends in the tutoring program, $\sum_{j \neq i} G_{ij} X_j$, due to the direct relationship between the latent attribute U_i and the friends of student i. Indeed, we previously established that:

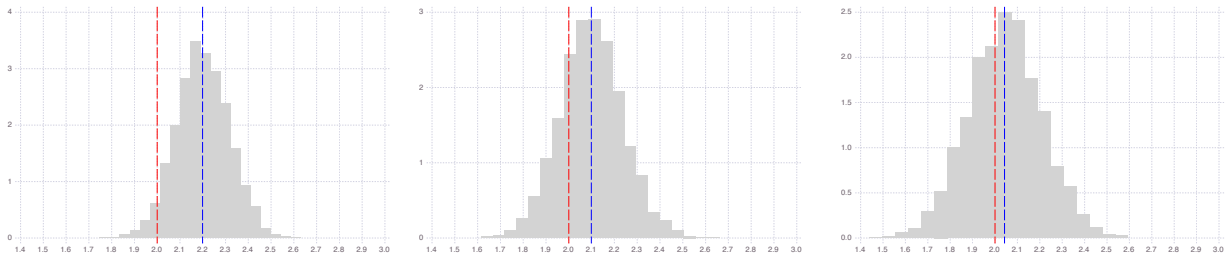
$$Cov\left(U_i, \sum_{j \neq i} G_{ij} X_j\right) = \mathbb{E}(X_1)\left(\mathbb{E}(U_i | D_{i+} > 0) - \mathbb{E}(U_i)\right) P(D_{i+} > 0)$$

suggesting an endogeneity bias in general if degree-centrality is indicative of sociability, i.e if $\mathbb{E}(U_i|D_{i+} > 0) \neq \mathbb{E}(U_i)$. Yet, a special edge case is if Assumption 11 holds which precludes isolated individuals and enforces that $Cov(U_i, \sum_{j \neq i} G_{ij} X_j) = 0$ (see Bramoullé et al. (2020) for a similar observation).

Assumption 11. *(No Isolated Individuals)*
$P(D_{i+} > 0) = 1$

Under such conditions, the fraction of treated friends becomes an exogenous regressor and under appropriate regularity conditions, the social interaction effects θ_0 can be consistently

estimated via least squares for cluster samples (Hansen and Lee (2019)). This identification result crucially hinges on Assumption 11 which is often explicitly or implicitly made in applied work but would be inadequate for any network dataset in which the fraction of isolated individuals is non-negligible: 23% in the Add Health dataset of Dieye and Fortin (2017) for example. Importantly, note that Assumption 11 is incompatible with CID models such as (3.2) given any finite set of individuals unless the graphon is degenerate; Assumption 10 excludes that possibility. Generally, we will have $\mathbb{E}(U_i|D_{i+} > 0) \neq \mathbb{E}(U_i)$ which in turn will lead to inconsistent estimates of the social effects as illustrated in Figure 3.1. The size of the bias will depend on the choice of the graphon, the distribution of latent attributes and importantly on the size of the network. The latter is illustrated in panels a, b and c of Figure 3.1 where the reduction in bias is clear as N increases. Note that this is little surprising since we are using a graphon-based model also suited to the modelling of dense network graphs. In Appendix 3.9.1, we show that standard methods to estimate Spatial Autoregressive models and the full linear in means (3.4) are also valid under Assumption 11.



(a) δ_{OLS} with $N = 20$ (b) δ_{OLS} with $N = 40$ (c) δ_{OLS} with $N = 80$

Figure 3.1: Contextual effect in the baseline *linear-in-means* model

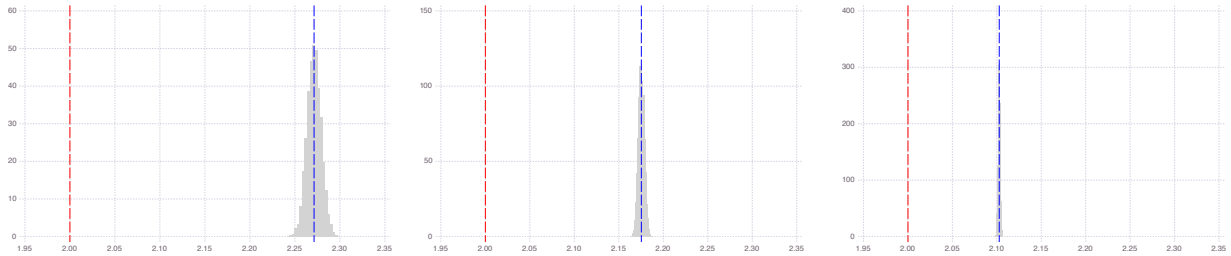
NOTES: The figures represent histograms of the least squares estimates for the contextual effect δ_0 in model (3.8) for $M = 5000$ Monte Carlo iterations. Each iteration considers $C = 200$ independent clusters with identical networks size $N = 20, 40, 80$. The link formation model is kept constant with $U, V \sim \mathcal{N}(0, 1)$ and $h(x, y) = \Phi\left(\frac{x+y}{\sqrt{2}}\right)$, where $\Phi(.)$ denote the CDF of a standard normal. Finally, the parameters of the outcome equation are $\alpha_0 = 0.2, \gamma_0 = 0.5, \delta_0 = 2.0$ and the covariates $X_i \sim$ Bernouilli($\frac{1}{2}$). The vertical dashed line in red indicates the true parameter value $\delta_0 = 2.0$ while the vertical dashed line in blue shows the average value of least squares estimates.

It is worth highlighting that the same logic would not apply in a model that considers a different weighting scheme for the terms capturing peer influence even under our Assumption 11, a fact first pointed out in Bramoullé et al. (2020). Indeed the above derivations leveraged both the absence of isolated individuals and the fact that G is row-normalized. In the

corresponding baseline *local-aggregate* model we have:

$$Y_i = \alpha_0 + \gamma_0 X_i + \delta_0 \sum_{j \neq i} D_{ij} X_j + U_i \tag{3.9}$$

If the graphon is increasing in its arguments, then the link probability is monotonic in the unobserved heterogeneity. Thus, popular agents, i.e agents with a high U_i will mechanically have more friends assigned to the tutoring program which implies $Cov(U_i, \sum_{j \neq i} D_{ij} X_j) > 0$ and in turn an upward bias for estimates of the exogenous effect. Interestingly, Figure 3.2, shows that the same CID process as in Figure 3.1 produces biases that are much more significant in the *local-aggregate* model, even for networks of relatively large size ($N = 80$ in panel c Figure 3.2).



(a) δ_{OLS} with $N = 20$ (b) δ_{OLS} with $N = 40$ (c) δ_{OLS} with $N = 80$

Figure 3.2: Contextual effect in the baseline *local-aggregate* model

NOTES: The figures represent histograms of the least squares estimates for the contextual effect δ_0 in model (3.8) for $M = 5000$ Monte Carlo iterations. Each iteration considers $C = 200$ independent clusters with identical networks size $N = 20, 40, 80$. The link formation model is kept constant with $U, V \sim \mathcal{N}(0, 1)$ and $h(x, y) = \Phi\left(\frac{x+y}{\sqrt{2}}\right)$, where $\Phi(.)$ denote the CDF of a standard normal. Finally, the parameters of the outcome equation are $\alpha_0 = 0.2, \gamma_0 = 0.5, \delta_0 = 2.0$ and the covariates $X_i \sim$ Bernouilli($\frac{1}{2}$). The vertical dashed line in red indicates the true parameter value $\delta_0 = 2.0$ while the vertical dashed line in blue shows the average value of least squares estimates.

Taking stock, these results show that standard models used in economics to analyze peer influence are subject to a potentially severe endogeneity bias when the self-selection of peers is not properly addressed. Under the condition of no isolated individuals, the baseline *linear-in-means* is unaffected but this restriction will be ruled out in several real cases of interest. Therefore, alternative strategies dealing with network endogeneity are necessary to identify and estimate the parameters of peer influence.

3.4 Identification with network endogeneity

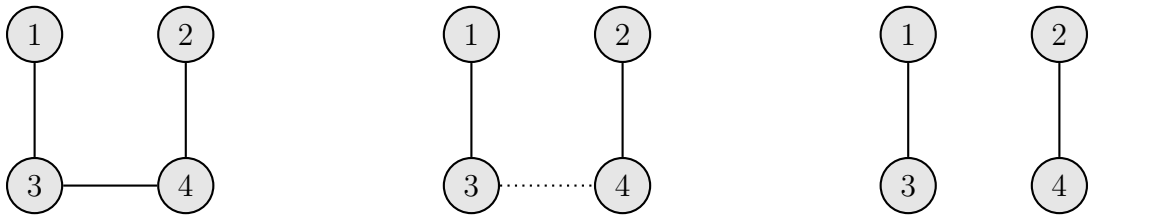
3.4.1 Symmetries of CID models

In this section, we present key symmetry properties of CID models that will help us address the problem of endogenous peer selection and motivate our identification and estimation strategies. The main result is that the latent attributes of any pair of automorphic nodes in a graph generated by a CID process are identically distributed conditional on the graph. To establish and clarify the meaning of this property, we require the introduction of a few graph-related concepts and it will be convenient to switch momentarily to the graph representation of a network instead of working through the adjacency matrix. That is, an undirected network of order N is a double $\mathcal{G} = (V(\mathcal{G}), E(\mathcal{G}))$, where $V(\mathcal{G}) = \{1, \ldots, N\}$ denotes the set of all nodes in the network and $E(\mathcal{G})$ is the set of edges across these nodes, i.e unordered pairs of vertices. For our purposes, it will also be useful to let $\bar{E}(\mathcal{G})$ record the set of non-edges so that we can equivalently represent $\mathcal{G}$ as the triplet $(V(\mathcal{G}), E(\mathcal{G}), \bar{E}(\mathcal{G}))$.

Definition 1. (Partial-edge subgraph). A partial-edge subgraph S of a graph $\mathcal{G}$, is a triplet $(V(S), E(S), \bar{E}(S))$ where $V(S) \subseteq V(\mathcal{G})$ is a subset of the vertices of G, $E(S) \subseteq E(\mathcal{G}) \cap V(S) \times V(S)$ is a subset of the edge set of $\mathcal{G}$, and $\bar{E}(S) \subseteq \bar{E}(\mathcal{G}) \cap V(S) \times V(S)$ is a subset of the non-edge set of G.

In the sequel, we will focus on the set of links of pairs of agents i and j in a network, i.e the i^{th} and j^{th} rows of the adjacency matrix, which constitutes an important example of a partial-edge subgraph. Figure 3.3 provides an illustration for the case of a tetrad network. In panel (b), the partial-edge subgraph S of network D depicted in panel (a) represents the friendships of agents 1 and 2 and is characterized by: $V(S) = \{1, 2, 3, 4\}$, $E(S) = \{13, 24\}$ and $\bar{E}(S) = \{12, 14, 23\}$. Importantly, note that S is silent about the connectivity of agents 3 and 4 as indicated by the dotted line between the two nodes in panel (b).

Figure 3.3: Subgraphs of a tetrad network



(a) Network D (b) Partial-edge subgraph $S = (D_1, D_2)$ of D (c) Partial subgraph $\tilde{S}$ of D

NOTES: A dotted line between two nodes indicates that their connectivity is unspecified by the subgraph.

As a result, it may generally be the case that: $\left|E(S)\right| + \left|\bar{E}(S)\right| \leq \binom{\left|V(S)\right|}{2}$. This feature of a partial-edge subgraph reflects the fact that in our leading estimation strategy, there will

be pairs of vertices whose connectivity we simply do not need to condition on.

We stress that our notion of a partial-edge subgraph is conceptually distinct from that of a partial subgraph (see Graham (2020)). The latter does not require "consistency" of the non edge set: $\bar{E}(\tilde{S}) \not\subseteq \bar{E}(\mathcal{G}) \cap V(\tilde{S}) \times V(\tilde{S})$ as illustrated in Panel (c) of figure 3.3 where $34 \in \bar{E}(\tilde{S})$ though agents 3 and 4 are originally connected in the network D of panel (a). Thus, a partial subgraph can modify the original topology while we can think of a partial-edge subgraph as providing a faithful but incomplete description of the underlying graph.

Definition 2. (Graph automorphism). Given a partial-edge subgraph $S = (V(S), E(S), \bar{E}(S))$ of a graph $\mathcal{G}$, a relabelling (i.e permutation) of the vertex set $\sigma : V(S) \mapsto V(S)$ is an automorphism of S if it maintains structure, that is if:

1. it preserves adjacency: $\forall (i,j) \in V(S), ij \in E(S) \implies \sigma(i)\sigma(j) \in E(S)$
2. it preserves non-adjacency: $\forall (i,j) \in V(S), ij \in \bar{E}(S) \implies \sigma(i)\sigma(j) \in \bar{E}(S)$

We let $Aut(S)$ denote the set of automorphisms of S.

With these definitions in hand, we can formally restate the assertion at the beginning of this section as follows: given a partial-edge subgraph S of an underlying graph $\mathcal{G}$ and $\sigma \in Aut(S)$, $U_i|S \sim U_{\sigma(i)}|S$. To show this result, we will prove the stronger fact that for any $A, B \subset V(S)$, $A \cap B = \emptyset$, such that $\sigma(A) = B, \sigma(B) = A$ for some $\sigma \in Aut(S)$, we have $U_A|S \sim U_B|S$. First, we must clarify the source of any potential symmetry, which of course lies in the exchangeability of CID models.

Lemma 14. *For any partial-edge subgraph $S = (V, E, \bar{E})$ and $\sigma \in Aut(S)$, we have* $P(S|U_V = u) = P\left(S|U_{\sigma(V)} = u\right)$

Proof.

$$
\begin{aligned}
P(S|U_V = u) &= \prod_{(i,j)\in E} h(u_i, u_j) \prod_{(i,j)\in \bar{E}} \left(1 - h(u_i, u_j)\right) \text{ (by Assumption 10)} \\
&= \prod_{(\sigma^{-1}(i),\sigma^{-1}(j))\in E} h(u_{\sigma(i)}, u_{\sigma(j)}) \prod_{(\sigma^{-1}(i),\sigma^{-1}(j))\in \bar{E}} \left(1 - h(u_{\sigma(i)}, u_{\sigma(j)})\right) \text{ (by def of } \sigma) \\
&= \prod_{(i,j)\in E} h(u_{\sigma(i)}, u_{\sigma(j)}) \prod_{(i,j)\in \bar{E}} \left(1 - h(u_{\sigma(i)}, u_{\sigma(j)})\right) \text{ (by def of } \sigma) \\
&= P\left(S|U_{\sigma(V)} = u\right)
\end{aligned}
$$

□

Now, Bayes theorem allows us to swap the conditioning variables to obtain the following corollary:

Corollary 14.1. *For any partial-edge subgraph $S = (V, E, \bar{E})$ and $\sigma \in Aut(S)$, we have $f_{U_V|S}(u|S) = f_{U_{\sigma(V)}|S}\left(u|S\right)$*

Proof.

$$\begin{aligned} f_{U_V|S}(u|S) &= \frac{P(S|U_V = u) f_{U_V}(u)}{P(S)} \\ &= \frac{P(S|U_V = u)) f_{U_{\sigma(V)}}(u)}{P(S)} \text{ (by Assumption 7)} \\ &= \frac{P(S|U_{\sigma(V)} = u) f_{U_{\sigma(V)}}(u)}{P(S)} \text{ (by Lemma 14)} \\ &= f_{U_{\sigma(V)}|S}(u|S) \end{aligned}$$

□

Corollary 14.1 shows a rather rich form of exchangeability in CID processes that we can now apply to prove the promised distributional identity:

Corollary 14.2. *For any partial-edge subgraph $S = (V, E, \bar{E})$ and $\sigma \in Aut(S)$, $\forall A, B \subset V$, $A \cap B = \emptyset$, such that $\sigma(A) = B, \sigma(B) = A$, we have $f_{U_A|S}(u|S) = f_{U_B|S}(u|S)$. In particular for $A = \{i\}$ and $B = \{\sigma(i)\}$, $i \in V$, we have $f_{U_i|S}(u|S) = f_{U_{\sigma(i)}|S}(u|S)$, i.e, automorphic nodes in S are identically distributed conditional on S.*

Proof.

$$\begin{aligned} f_{U_A,U_B|S}(u_a, u_b) &= \int_{\mathcal{U}^{|V|-|A|-|B|}} f_{U_A,U_B,U_{V\setminus A\cup B}|S}(u_a, u_b, u_{V\setminus A\cup B}) du_{V\setminus A\cup B} \\ &= \int_{\mathcal{U}^{|V|-|A|-|B|}} f_{U_\sigma(A),U_\sigma(B),U_{\sigma(V\setminus A\cup B)}|S}(u_a, u_b, u_{V\setminus A\cup B}) du_{V\setminus A\cup B} \\ &\quad \text{(by Corollary 14.1)} \\ &= \int_{\mathcal{U}^{|V|-|A|-|B|}} f_{U_B,U_A,U_{\sigma(V\setminus A\cup B)}|S}(u_a, u_b, u_{V\setminus A\cup B}) du_{V\setminus A\cup B} \\ &= \int_{\sigma(\mathcal{U}^{|S|-|A|-|B|})^{-1}} f_{U_B,U_A,U_{V\setminus A\cup B}|S}(u_a, u_b, u_{\sigma(V\setminus A\cup B)^{-1}}) du_{\sigma(V\setminus A\cup B)^{-1}} \\ &= \int_{\mathcal{U}^{|V|-|A|-|B|}} f_{U_B,U_A,U_{V\setminus A\cup B}|S}(u_a, u_b, u_{V\setminus A\cup B}) du_{V\setminus A\cup B} \\ &= f_{U_B,U_A|S}(u_a, u_b) \end{aligned}$$

We conclude that U_A and U_B are exchangeable conditional on S and consequently identically distributed conditional on S. □

Corollary 14.2 is a simple, yet powerful exchangeability result that will be essential to deal with network endogeneity and identify the social effects θ_0 . We defer the discussion of how to leverage this result in linear social interactions models to section 3.4.2.2.

3.4.2 The baseline model with contextual effect

3.4.2.1 Observed control function: adjusting for degree centrality

This section introduces a simple observed control function procedure to deal with network endogeneity that we present in the baseline model that only features exogenous effects. The stragegy is a natural starting point and though it is considerably limited, we will see that it carries valuable insights on how to approach the problem of endogenous peer selection more systematically (Section 3.4.2.2). Consider the following specification

$$Y_i = \alpha_0 + \gamma_0 X_i + \delta_0 \sum_{j \neq i} \omega_{ij}(D_i) X_j + U_i \tag{3.10}$$

Recall that $\omega_{ij}(D_i) = G_{ij}$ for the *linear-in-means* and $\omega_{ij}(D_i) = D_{ij}$ for the *local-aggregate* model. In light of Assumption 8, we know that X_i is an exogenous covariate but the characteristics of friends $\sum_{j \neq i} \omega_{ij}(D_i) X_j$ are generally not due to the endogeneity of the peer group: U_i and D_i are correlated. One possible solution to address this problem is to proceed as follows. By Assumptions 7,8 and 10 we have: $\mathbb{E}[U_i|X, D_i] = \mathbb{E}[U_i|D_i]$. Furthermore, since D_{ij} are dummy variables, we recognize a saturated regression model entailing that $\mathbb{E}[U_i|D_i]$ is a polynomial in the D_{ij}. For example, with a triad network, $N = 3$, we can write:

$$\mathbb{E}[U_1|D_{12}, D_{13}] = a + bD_{12} + cD_{13} + dD_{12}D_{13}$$

In general, this decomposition adds $2^{(N-1)}$ parameters to estimate but this can be reduced in our context by exploiting a symmetry property of the network formation model (3.2) referenced in Lemma 15. This useful feature harks back to the work of Altonji and Matzkin (2005) who studied the implications of exchangeability for identification in nonseparable panel data settings with endogenous regressors.

Lemma 15 (Permutation Invariance)**.** *The conditional density of* $U_i|D_i = d_i$ *is invariant to permutations in* d_i*.*

$$f_{U_i|D_i}(u_i|d_{i1}, \ldots, d_{i(j-1)}, d_{i(j+1)}, \ldots, d_{iN}) = f_{U_i|D_i}(u_i|d_{i\sigma(1)}, \ldots, d_{i\sigma(j-1)}, d_{i\sigma(j+1)}, \ldots, d_{i\sigma(N)})$$

$\forall \sigma$ *a permutation of* $\{1, \ldots, N\} \setminus \{i\}$

By definition:

$$\mathbb{E}[U_1|D_{12} = d_{12}, D_{13} = d_{13}, \ldots, D_{1N} = d_{1N}] = \int u f_{U_1|D_{12}, D_{13}, \ldots, D_{1N}}(u|d_{12}, d_{13}, \ldots, d_{1N}) du$$

and since $f_{U_1|D_1}(u|d_1)$ is permutation invariant in d_1 by Lemma 15, it follows that $\mathbb{E}[U_1|D_1 = d_1]$ is a symmetric polynomial in $d_1 = (d_{12}, d_{13}, \ldots, d_{1N})$. Returning to the example of 3 agents, this symmetry implies:

$$\begin{aligned}\mathbb{E}[U_1|D_{12}, D_{13}] &= a + bD_{12} + cD_{13} + dD_{12}D_{13} \\ &= \mathbb{E}[U_1|D_{13}, D_{12}] = a + bD_{13} + cD_{12} + dD_{13}D_{12}\end{aligned}$$

this entails $b = c$, so:

$$\mathbb{E}[U_1|D_{12}, D_{13}] = a + b(D_{12} + D_{13}) + dD_{12}D_{13}$$

which leaves us with only N nuisance parameters instead of $2^{(N-1)}$. Finally, since $\mathbb{E}[U_i|D_i]$ is a symmetric polynomial, we can succinctly express it in terms of degree centrality of the agent. In the simple case of a triad, we have:

$$\begin{aligned}\mathbb{E}[U_1|D_{12}, D_{13}] &= a + b(D_{12} + D_{13}) + dD_{12}D_{13} \\ &= a + bD_{1+} + dD_{12}D_{13} \\ &= a + (b - \frac{1}{2}d)D_{1+} + \frac{d}{2}D_{1+}^2\end{aligned}$$

And more generally with N agents, it is not difficult to see that we will obtain:

$$\mathbb{E}[U_i|D_i] = \sum_{k=0}^{N-1} c_{k,N} D_{i+}^k$$

Intuitively, this expression says that the degree centrality of an agent acts as a "sufficient statistic" to approximate it's unobserved heterogeneity. We can write:

$$\begin{aligned}Y_i &= \alpha_0 + \gamma_0 X_i + \delta_0 \sum_{j \neq i} \omega_{ij}(D_i) X_j + U_i \\ &= \alpha_0 + \gamma_0 X_i + \delta_0 \sum_{j \neq i} \omega_{ij}(D_i) X_j + \mathbb{E}[U_i|X, D_i] + (U_i - \mathbb{E}[U_i|X, D_i]) \\ &= \alpha_0 + \gamma_0 X_i + \delta_0 \sum_{j \neq i} \omega_{ij}(D_i) X_j + \sum_{k=0}^{N-1} c_{k,N} D_{i+}^k + V_i\end{aligned}$$

with $V_i = U_i - \mathbb{E}[U_i|X, D_i]$, the reduced-form error satisfying $\mathbb{E}[V_i|X, D_i] = 0$ by construction. Under this alternative formulation of equation (3.10) that controls for the degree centrality of the agent - an observable quantity for the econometrician - $\sum_{j \neq i} \omega_{ij}(D_i) X_j$ is no longer endogenous. Therefore, under standard rank conditions, the identification and consistent estimation of the social effects becomes possible when the data consists of a collection of networks of the same size and generated by the same DGP. These strong conditions are necessary to ensure that the coefficients of the control function: $(c_{k,N})_{k=0}^N$ are identical across networks. One caveat of course is that in practice the researcher is likely to observe networks

of varying sizes which invalidates this approach.[5] Another weakness of this method is that it requires the distribution of latent attributes, the distribution of dyadic schocks and the graphon to be identical across networks. We introduce the *degree-matching* approach in the ensuing section to overcome some of these limitations.

3.4.2.2 The degree-matching approach

The reduced-form expression of specification (3.10) featuring the observed control function suggests taking pairwise differences of agents having identical degree centrality to eliminate the network fixed effect and the counfounding effect of latent attributes; a strategy reminiscent of Honoré and Powell (1994), Aradillas-Lopez et al. (2007). We refer to this approach as *degree-matching*. The logic is simple: since degree centrality approximates the latent individual characteristic U_i well, and since it is observable, degree-matching should approximately remove these sources of endogeneity from the estimating equations. This is similar in spirit to how first-differencing removes individual heterogeneity in a linear panel data setting. To illustrate, let us consider an extended version of the baseline model with both network fixed effect and latent exogenous effect:

$$Y_i = \alpha_0 + \gamma_0 X_i + \delta_0 \sum_{j \neq i} \omega_{ij}(D_i) X_j + \epsilon_i$$
$$\epsilon_i = A_D + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i = U_i + \lambda_0 \sum_{j \neq i} \omega_{ij}(D_i) U_j$$

Naturally, the added complexity of the error term in the outcome equation exacerbates the problem of peer group endogeneity. The intuition described above for the validity of degree-matching as an identification strategy can be formalized through the conditional moment restrictions of Theorem 6. Before stating its content, we need a few more exchangeability results that we discuss at length below.

Corollary 15.1. *Consider N agents. $\forall (d_i, d_j) \in \{0,1\}^{N-1} : d_{i+} = d_{j+}$, we have:*

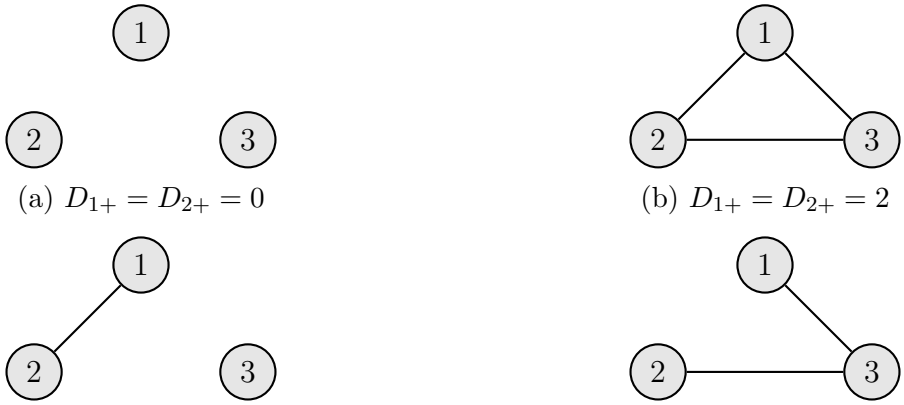
1. $U_i | D_i = d_i, D_j = d_j \sim U_j | D_i = d_i, D_j = d_j$
2. $\forall (k,l) \in \{1, \ldots, N\} \setminus \{i,j\}$:
   - *a)* $d_{ik} = 1,\ d_{il} = 0,\ d_{jk} = 0,\ d_{jl} = 1 \implies U_k | D_i = d_i, D_j = d_j \sim U_l | D_i = d_i, D_j = d_j$
   - *b)* $d_{ik} = 1,\ d_{il} = 1,\ d_{jk} = 0,\ d_{jl} = 0 \implies U_k | D_i = d_i, D_j = d_j \sim U_l | D_i = d_i, D_j = d_j$

[5] Similar assumptions have nonetheless antecedents in the literature: Moffitt et al. (2001) discusses versions of the *linear-in-means* model where the researcher observes multiple groups of the same size

Corollary 15.1 is merely an implication of Corollary 14.2 and follows from noticing that $S = \{D_i = d_i, D_j = d_j\}$ is a partial-edge subgraph of network D, and that the restriction $d_{i+} = d_{j+}$ imposes that: 1) i and j are automorphic nodes in S, 2) the "exclusive friends" of i and j are exchangeable within and across groups.
The conclusions can be easily grasped in the case of a triad network. Consider for example the event in which agent 1 and agent 2 have identical degree centrality when $N = 3$:

$$\{D_{1+} = D_{2+}\} = \{D_{12} = 0, D_{13} = 0, D_{23} = 0\} \cup \{D_{12} = 1, D_{13} = 0, D_{23} = 0\} \\ \cup \{D_{12} = 0, D_{13} = 1, D_{23} = 1\} \cup \{D_{12} = 1, D_{13} = 1, D_{23} = 1\}$$

Figure 3.4: Triads with $\{D_{1+} = D_{2+}\}$



(a) $D_{1+} = D_{2+} = 0$ (b) $D_{1+} = D_{2+} = 2$

(c) $D_{1+} = D_{2+} = 1$

NOTES: This figure depicts all triad configurations consistent with the event $\{D_{1+} = D_{2+}\}$.

Figure 3.4 depicts all four triad configurations consistent with the latter. Notice the evident symmetry of each subgraph with respect to vertices 1 and 2. Since the U_i's and V_{ij}'s are iid, in the absence of node specific covariates, node labels are meaningless so agent 1 and agent 2 are exchangeable in each triad configuration of Figure 3.4. This observation implies part 1 of Corollary 15.1; part 2 is equally intuitive.

Importantly, note that the distributional statements of Corollary 15.1 are conditional on the observed links of two agents only and not on the full network of interactions as the results would not generally hold otherwise. To understand why, consider the pentad wiring displayed in Figure 3.5. There, agent 1 and agent 2 both have degree 1 but are linked to agents that differ in gregariousness: agent 1 is friend with agent 3 with degree 3 while agent 2 is friend with agent 5 with degree 2. Thus, they cannot be permuted in contrast to agent 1 and agent 4 who are automorphic. Intuitively, the level of popularity of an individual is not solely reflected by her number of friends but also by her position in the network. From a mathematical vantage point, the reason is that with a CID model such as (3.2), indirect connections $(D_k)_{k \neq i,j}$ are informative for U_i and U_j because D_{il} and D_{jl} are correlated to D_{kl}

through U_l. As a result, in this example, with d representing the pentad wiring of Figure 3.5, we have $U_1|D_1 = d_1, D_2 = d_2 \sim U_2|D_1 = d_1, D_2 = d_2$ but $U_1|D = d \not\sim U_2|D = d$.

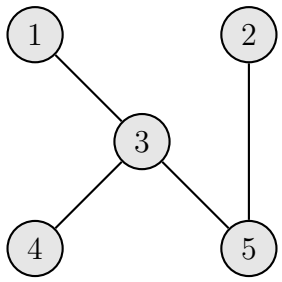


Figure 3.5: Asymmetric pentad wiring relative to nodes 1 and 2

Theorem 6. *Let $g(.)$ be a measurable function of (D_i, D_j, X) such that $\mathbb{E}\left[\left|g(D_i, D_j, X)\right|^2\right] < \infty$. Then*

$$\mathbb{E}\left[g(D_i, D_j, X)(\epsilon_i - \epsilon_j)|X, D_{i+} = D_{j+}\right] = 0 \tag{3.11}$$

The proof of Theorem 6 follows from two observations. First, for a network of order N, the law of total expectations in conjunction with Assumptions 7,8,10, allow us to express the moment condition (3.11) as:

$$\mathbb{E}\left[g(D_i, D_j, X)(\epsilon_i - \epsilon_j)|X, D_{i+} = D_{j+}\right] = \sum_{(d_i,d_j):d_{i+}=d_{j+}} \frac{P(D_i = d_i, D_j = d_j)}{P(D_{i+} = D_{j+})} g(d_i, d_j, X) \times \mathbb{E}\left[\tilde{\epsilon}_i - \tilde{\epsilon}_j|D_i = d_i, D_j = d_j\right]$$

Second, $\tilde{\epsilon}_i - \tilde{\epsilon}_j$ involves two types of quantities: $U_i - U_j$ and $U_k - U_l$ where k and l are exclusive friends of i and j respectively. Therefore, by Corollary 15.1, it follows that conditional on $D_i = d_i, D_j = d_j$ where $d_{i+} = d_{j+}$, $\mathbb{E}[\tilde{\epsilon}_i - \tilde{\epsilon}_j|D_i = d_i, D_j = d_j] = 0$.

Theorem 6 is a very useful result for linear social interaction models; it says that by matching individuals on degree centrality we have the guarantee that any statistical characteristic involving their links and the covariates of interest X is exogenous. In particular, this is true for the difference in friends characteristics by setting $g(D_i, D_j, X) = \sum_{k\neq i} \omega_{ik}(D_i)X_k - \sum_{l\neq j} \omega_{jl}(D_j)X_l$.

To operationalize the degree-matching idea, let $Z_i = (X_i, \sum_{j\neq i} \omega_{ij}(D_i)X_j)'$, $i = 1, \ldots, N$, be the vector of explanatory variables and consider the objective function:

$$\mathcal{Q}(\theta) = \mathbb{E}\left[\left(Y_i - Y_j - (Z_i - Z_j)'\theta\right)^2 |D_{i+} = D_{j+}\right], \quad \theta \in \mathbb{R}^2$$

Additionally, suppose that the following rank condition for the regressors holds:

Assumption 12. *(Rank condition)*
$\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)'|D_{i+} = D_{j+}\right]$ *is non singular*

Assumption 12 is an identification condition analogous to the standard full rank assumption on the regressors in linear regression models. It is implied by Assumption 7-10 when $K = 1$.

Lemma 16. *In the baseline linear-in-sums and linear-in-means model with* $dim(X_i) = K = 1$*, Assumption 7-10 implies Assumption 12.*

Then, the following result holds:

Proposition 6. *Suppose Assumptions 7-10 and 12 hold. Then* $\mathcal{Q}(\theta)$ *is uniquely minimized at* θ_0*.*

Proposition 6 shows formally that matching agents on degree centrality is a fruitful approach to identify the social effects θ_0. Conceptually, the method first deals with network endogeneity through matching and then exploits covariate variation in subgraphs with identical degree sequence to identify θ_0 in the same way that variation in group size in a setting where individuals interact in groups helps identify the social parameters (Davezies et al. (2009)). A convenient by-product of this approach is that it automatically deals with the problem of correlated effects. Heuristically, when $N = 2$, identification of the social parameters via *degree-matching* comes down to comparing the reduced-form coefficients for connected and disconnected pairs and a similar logic applies when $N = 3$. For larger networks $N \geq 4$, the identification of θ_0 is helped by the presence of network wirings in which two agents have identical degree centrality but do not share exactly the same set of friends. Such network configurations, at least in the case $K = 1$, guarantee that Assumption 6 is verified (see Lemma 16).

Remark 11.
In light of the results of section 3.4.1, another strategy to identify the social effects in the same vein as *degree-matching* would be to match symmetric agents in the full network: *orbit-matching*. Consider the shorthand, $i \sim_G j$, to denote two nodes of a graph $\mathcal{G}$ that are automorphic and let $\mathcal{O}_{\mathcal{G}}(i) = \{j \sim_G i, j \in V(\mathcal{G})\} = \{\sigma(i)|\sigma \in Aut(\mathcal{G})\}$ denote the $\mathcal{G}$-orbit of i. Orbits partition the vertex set into disjoint equivalence classes. Then, it is natural to consider the criterion function:

$$\mathcal{S}(\theta) = \mathbb{E}\left[\left(Y_i - Y_j - (Z_i - Z_j)'\theta\right)^2 |j \in \mathcal{O}_{\mathcal{G}}(i)\right], \quad \theta \in \mathbb{R}^2$$

$\mathcal{S}(\theta)$ and $\mathcal{Q}(\theta)$ are similar and in fact coincide when $N \leq 3$. In general however, the two criteria will be different as $i \approx j \implies D_{i+} = D_{j+}$ but
$D_{i+} = D_{j+} \not\implies i \approx j$. The pentad wiring of Figure 3.5 is an illustration of the latter. In practice, from an estimation perspective, this approach is likely to be less tractable than *degree-matching* as it will require to first determine the automorphism group of $\mathcal{G}$ to characterize the $\mathcal{G}$-orbits and make the pairwise difference approach feasible. Generally, this initial

step is a non trivial task called the "graph isomorphism problem" and is a well known computational problem in discrete mathematics. However, in our setting of many small networks, the computational considerations are likely to be less important which makes *orbit-matching* a potentially viable strategy [6]. I leave the investigation of this idea for future work.

To conclude this section, let us highlight that *degree-matching* also presents some drawbacks: its inability to identify the "own-effect" γ_0 when the dependence between X_i and U_i is unrestricted - a general limitation of pairwise difference procedures (Aradillas-Lopez et al. (2007)). In Section 3.5, we show that while it is possible to relax the independence assumption between X_i and U_i, salvaging the identification of the entire parameter vector θ_0 requires a type of exclusion restriction.

3.4.3 SAR and the full model

Extending the *degree-matching* approach to identify the social effects in Spatial Autoregressive (SAR) models is relatively straightforward. The SAR model, popular in spatial econometrics posits the following relation:

$$Y_i = \alpha_0 + \beta_0 \sum_{j \neq i} \omega_{ij}(D_i) Y_j + \gamma_0 X_i + \epsilon_i \tag{3.12}$$

$$\epsilon_i = A_D + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i = U_i + \lambda_0 \sum_{j \neq i} \omega_{ij}(D_i) U_j \tag{3.13}$$

It is obtained by setting $\delta_0 = 0$ in specification (3.7). In other words, the model assumes away (observable) exogenous effects. Commonly, $\lambda_0 = 0$ but in line with our previous discussions, we will consider the possibility of latent exogenous effect: $\lambda_0 \neq 0$. This constitutes a system of simultaneous equations which even in the absence of endogenous peer selection would pose an endogeneity problem due to the correlation between the outcome of peers and the error term, i.e the reflection problem (Manski (1993)). The traditional approach to identify and estimate $\theta_0 = (\beta_0, \gamma_0)'$ without correlated effects when the covariates and the network are exogenous is to instrument peer response $\sum_{j \neq i} \omega_{ij}(D_i) Y_j$ by the characteristics of peers $\sum_{j \neq i} \omega_{ij}(D_i) X_j$. Of course, this is not directly applicable in our setting with endogenous peers as $\sum_{j \neq i} \omega_{ij}(D_i) X_j$ is correlated with the latent individual attributes and potentially with the network fixed effect. To identify the social effects, we propose combining the IV approach with the *degree-matching* procedure discussed previously for the baseline models. A heuristic rationale is that since *degree-matching* jointly deals with network endogeneity and correlated effects; upon matching individuals by degree centrality we are free to use the remaining exogenous variation from the instrument to solve the reflection problem and identify θ_0.

Let $Z_i = (X_i, \sum_{j \neq i} \omega_{ij}(D_i) Y_j)'$ denote the vector of regressors and

[6] There are efficient algorithms such as Nauty (No AUTomorphisms, Yes?) that can compute the automorphism group of graphs with less than 100 nodes in under a second.

$W_i = (X_i, \sum_{j \neq i} \omega_{ij}(D_i)X_j)'$ denote the vector of instruments. Under *degree-matching* the instruments satisfy the exclusion restriction:

$$\begin{aligned}\mathbb{E}\left[(W_i - W_j)(\epsilon_i - \epsilon_j)|D_{i+} = D_{j+}\right] &= \mathbb{E}\left[\underbrace{\mathbb{E}\left[(W_i - W_j)(\epsilon_i - \epsilon_j)|X, D_{i+} = D_{j+}\right]}_{=0 \text{ by Theorem 6}}|D_{i+} = D_{j+}\right] \\ &= 0\end{aligned}$$

which combined with Assumption 13 hereinafter can be leveraged as a basis for the identification of θ_0.

Assumption 13. *(Instrument relevance)*
$\mathbb{E}\left[(W_i - W_j)(Z_i - Z_j)'|D_{i+} = D_{j+}\right]$ *is full rank*

Indeed, it follows that:

$$\theta_0 = \mathbb{E}\left[(W_i - W_j)(Z_i - Z_j)'|D_{i+} = D_{j+}\right]^{-1}\mathbb{E}\left[(W_i - W_j)(Y_i - Y_j)|D_{i+} = D_{j+}\right]$$

The treatment of the full linear social interaction model is more challenging and it will be useful to cover the *linear-in-means* and the *linear-in-sums* separately. Let us begin with the *linear-in-means* specification (3.4) assuming temporarily the absence of network fixed effect. Rewriting the equations in matrix form, we have:

$$\begin{aligned}Y &= \alpha_0\iota + \beta_0 GY + \gamma_0 X + \delta_0 GX + U + \lambda_0 GU \\ &\implies (I - \beta_0 G)Y = \alpha_0\iota + (\gamma_0 I + \delta_0 G)X + \underbrace{(I + \lambda_0 G)U}_{=\epsilon}\end{aligned}$$

where ι is a conformable vector of ones. Given our primitive assumption that $|\beta_0| < 1$, the matrix $(I - \beta_0 G)$ is diagonally dominant, thus invertible with $(I - \beta_0 G)^{-1} = \sum_{k=0}^{\infty} \beta_0^k G^k$. Therefore:

$$\begin{aligned}Y &= \alpha_0(I - \beta_0 G)^{-1}\iota + (\gamma_0 I + \delta_0 G)X + (I - \beta_0 G)^{-1}(I + \lambda_0 G)U \\ &= \alpha_0(I - \beta_0 G)^{-1}\iota + \gamma_0 X + (\gamma_0\beta_0 + \delta_0)\sum_{k=0}^{\infty}\beta_0^k G^{k+1}X + (I - \beta_0 G)^{-1}(I + \lambda_0 G)U\end{aligned}$$

and hence

$$GY = \alpha_0 G(I - \beta_0 G)^{-1}\iota + \gamma_0 GX + (\gamma_0\beta_0 + \delta_0)\sum_{k=0}^{\infty}\beta_0^k G^{k+2}X + G(I - \beta_0 G)^{-1}(I + \lambda_0 G)U$$

This last expression suggests that under the usual assumption of network exogeneity: $\mathbb{E}[U|X, G] = 0$, and provided that $(\gamma_0\beta_0 + \delta_0) \neq 0$, G^2X, $G^3X \ldots$ may be used as instruments

for GY (Bramoullé et al. (2009)). However, in our model, the exclusion restrictions for the conventional instruments generally fail as:

$$\forall k \geq 0, \quad \mathbb{E}(G^{k+2}G(I-\beta_0 G)^{-1}(I+\lambda_0 G)U) \neq 0$$

Thus an alternative methodology is required. In the specific case of the *linear-in-means* specification, we can show that as in SAR models, IV-*degree-matching* offers a potential solution. This is motivated by the following moment conditions:

Lemma 17. $\forall m \in \mathbb{N}, \quad \mathbb{E}\left[\left((G^m X)_i - (G^m X)_j\right)(\epsilon_i - \epsilon_j)\middle| D_{i+} = D_{j+}\right] = 0$

Lemma 17 shows in particular that the friends' friends' average characteristics, the baseline instrument under network exogeneity remains viable after matching individuals by degree centrality. Interestingly, this result hinges again on the fact that the matrix G is row-normalized and thus does not carry over to the *local-aggregate* model. Note that the result is also unaffected by the presence of network fixed effects. Utimately, the validity of this strategy will rest on the relevance of the set of instruments which may be problematic depending on network size in contrast to the simpler models discussed previously. To illustrate this point, let $Z_i = (X_i, (GX)_i, (GY)_i)'$ denote the vector of covariates and $W_i = (X_i, (GX)_i, (G^2X)_i)'$ its corresponding vector of instruments. In a population of networks of size $N = 2$, we will not be able to pin down the social effects having more parameters than equations per cluster. For $N = 3$, the set of network wirings such that two individuals have the same degree centrality (see Figure 3.4) exhibit such symmetry that the friends' friends' average characteristics will be colinear to own characteristics and friends' average characteristic. Similar difficulties occur for the case $N = 4, 5$. Starting at $N \geq 6$, there are network configurations where two individuals have the same degree centrality and have friends whose friends do not fully overlap. Hence, there will be variation in $(G^2X)_i - (G^2X)_j$ unrelated to that in $X_i - X_j$ and $(GX)_i - (GX)_j$. In turn, this means that in most networks of modest size the combination of IV and *degree-matching* in *linear-in-means* models provides a way to pin down the social effects.

Unfortunately, as hinted above, there is no equivalent of Lemma 17 for the *local-aggregate* model although, the model has a similar reduced form motivating the use of DX, D^2X,... as instruments under network exogeneity (Liu and Lee (2010)). The underlying reason is that for $m \geq 2$, $(D^m X)_i - (D^m X)_j$ involves links outside the partial-edge subgraph $\{D_i, D_j\}$ that must be integrated over in quantities such as: $\mathbb{E}\left[\left((D^m X)_i - (D^m X)_j\right)(\epsilon_i - \epsilon_j)\middle| D_{i+} = D_{j+}\right]$. Integrating the product of $(D^m X)_i - (D^m X)_j$ and $\epsilon_i - \epsilon_j$ over the linking decisions of agents different from i and j necessarily involves computations of this integrand over network configurations that are asymmetric with respect to i and j - even though they have the same degree centrality - which in turn prevents the use of an exchangeability argument in the vein of Corollary 14.2 to obtain an orthogonality condition. The fact that the *linear-in-means* model is immune to this issue is an artificial product of the row-normalization of

the adjacency matrix. As a concluding remark, note that this issue would not arise with *orbit-matching* as the conditioning event $\{j \in \mathcal{O}_{\mathcal{G}}(i)\}$ requires a symmetry over the entire graph in contrast to $\{D_{i+} = D_{j+}\}$.

3.5 Adding homophily on observable characteristics

We now briefly discuss how to generalize our identification strategy to the case of a CID model that accommodates homophily on observables. Let R_i denote an observable individual attribute distinct from X_i entering the following link formation process:

$$D_{ij} = \mathbb{1}\left\{h(U_i, R_i, U_j, R_j) \geq V_{ij}\right\} \mathbb{1}\left\{i \neq j\right\} \tag{3.14}$$

with the graphon $h(.)$ symmetric in $(U_i, R_i), (U_j, R_j)$. For simplicity, we will focus on the case where R_i is a discrete random variable with finite support $\mathcal{R} = \{r_1, \ldots, r_L\}$. Nothing that follows essentially hinges upon this restriction [7]. These observable attributes partition the population into $|\mathcal{R}| = L$ categories that we will call types. Here, we relax the assumption of independence between U_i and X_i and assume instead that the following distributional exclusion restriction (Powell (1994), p. 2484)) holds:

Assumption 14. *(Distributional exclusion restriction)*
$f_{U_i|X_i,R_i}(u_i|x_i, r_i) = f_{U_i|R_i}(u_i|r_i)$

with the joint distribution of (X_i, R_i) left unrestricted. This kind of assumption is familiar from the work of Blundell and Powell (2004) on semiparametric binary response models with endogenous regressors. In the present context, it corresponds to a redundancy condition: X_i cannot have any predictive power over U_i conditional on R_i. Then, a straightforward adaptation of the previous methodology is to match agents on: degree centrality, their type, and the types of their friends. The intuition for that is similar to before: if agent i and agent j are of the same type and have an identical number of friends of the same type, then conditional on this sole information, agent i and agent j are exchangeable (in a partial-edge graph sense). A formal proof is omitted for brevity but would be a straightforward variant of Corollary 15.1.

Following this logic, the identification of θ_0 in, say, the baseline model could be established from the criterion function:

$$\mathcal{Q}(\theta) = \mathbb{E}\left[\left(Y_i - Y_j - (Z_i - Z_j)'\theta\right)^2 | D_{i+} = D_{j+}, R_i = R_j, R_{i(-j)} = R_{j(-i)}\right]$$

where we use the shorthand $R_{i(-j)} = \{r_k \in \mathcal{R} | k \in \{1, \ldots, N\} \setminus \{j\} : d_{ik} = 1\}$, i.e the types of i's friends that are not j. The form of $\mathcal{Q}(\theta)$ makes it clear that in general any component of X_i that overlaps with R_i will be wiped out in the objective function which in turn compromises the identification of the associated parameters. The role of Assumption 14 is precisely to guarantee that there remains identifying varation from X_i after the matching step to pin down θ_0.

[7] The continous case can be tackled with an appropriate choice of bandwidth and kernel.

3.6 Estimation

3.6.1 Baseline models

For ease of illustration, we start by presenting the estimators based on the observed control function approach and the *degree-matching* strategy for the baseline peer effect models featuring only exogenous effects, i.e equations (3.10).

Following Section 3.2.3, we consider a collection of C independent networks with size $N_1, \dots, N_C$ respectively. Let $n = \sum_{c=1}^{C} n_c$, where $n_c = \binom{N_c}{2}$ corresponds to the number of unique dyads in cluster c. As discussed in Section 3.2.3, although we allow networks to be of varying sizes, it is assumed that friendships are generated according to the same CID model (3.2) and that the same outcome equation (3.10) applies to each cluster. In this setting, the *degree-matching* estimator $\hat{\theta}_{DM}$ for our baseline peer effect model takes the form:

$$\hat{\theta}_{DM} = \arg\min_{\theta} \frac{1}{n} \sum_{c=1}^{C} \sum_{i=1}^{N_c-1} \sum_{j=i+1}^{N_c} \mathbb{1}\{D^c_{i+} = D^c_{j+}\} \left(Y^c_i - Y^c_j - (Z^c_i - Z^c_j)'\theta \right)^2$$

with a closed form solution conveniently given by: $\hat{\theta}_{DM} = \hat{Q}_n^{-1} \hat{S}_n$ with

$$\hat{Q}_n = \frac{1}{n} \sum_{c=1}^{C} \sum_{i=1}^{N_c-1} \sum_{j=i+1}^{N_c} \mathbb{1}\{D^c_{i+} = D^c_{j+}\} \Delta Z_{cij} \Delta Z'_{cij}$$

$$\hat{S}_n = \frac{1}{n} \sum_{c=1}^{C} \sum_{i=1}^{N_c-1} \sum_{j=i+1}^{N_c} \mathbb{1}\{D^c_{i+} = D^c_{j+}\} \Delta Y_{cij} \Delta Z_{cij}$$

$$\text{where } \Delta \xi_{cij} = \xi^c_i - \xi^c_j$$

where Z_i denotes the vector of explanatory variables: $Z_i = (X_i, \sum_{j \neq i} D_{ij} X_j)'$ for the *local-aggregate* model and $Z_i = (X_i, \sum_{j \neq i} G_{ij} X_j)'$ for the *linear-in-means*. Observe the normalization by n here rationalized by the fact that each cluster c contributes precisely a total of n_c unique pairs of observations.

We draw on the work of Hansen and Lee (2019) on asymptotic theory for clustered samples to prove the consistency and asymptotic normality of our estimators. It is helpful to express $\hat{Q}_n$ and $\hat{S}_n$ in matrix form to establish a clear connection to their work. To that end, we introduce the boldface indices $\mathbf{i} = \mathbf{1}, \mathbf{2}, \dots$ as an index for dyads in each cluster $c = 1, \dots, C$ and abusing notations, we let $\mathbf{i}$ also denote the set $\{i_1, i_2\}$ where i_1 and i_2 are the agents comprising the dyad $\mathbf{i}$. With these notations in hand, let us introduce:
$\widetilde{\Delta} Z_{c\mathbf{i}} = \mathbb{1}\{D^c_{i_1+} = D^c_{i_2+}\} \Delta Z_{ci_1i_2}$ the $2K \times 1$ vector of individual regressors and
$\widetilde{\Delta} Z_c = (\widetilde{\Delta} Z_{c\mathbf{1}}, \dots, \widetilde{\Delta} Z_{c\mathbf{n_c}})'$ the $n_c \times 2K$ matrix of regressors for the cth network. Analogously, let $\widetilde{\Delta} Y_c = (\widetilde{\Delta} Y_{c\mathbf{1}}, \dots, \widetilde{\Delta} Y_{c\mathbf{n_c}})'$, $\widetilde{\Delta} U_c = (\widetilde{\Delta} U_{c\mathbf{1}}, \dots, \widetilde{\Delta} U_{c\mathbf{n_c}})'$ be the $n_c \times 1$ matrix of outcomes, respectively errors in the cth network, with $\widetilde{\Delta} Y_{c\mathbf{i}} = \mathbb{1}\{D^c_{i_1+} = D^c_{i_2+}\} \Delta Y_{ci_1i_2}$ and

$\widetilde{\Delta}U_{c\mathbf{i}} = \mathbb{1}\{D^c_{i_1+} = D^c_{i_2+}\}\Delta U_{ci_1i_2}$. Then, we can write:

$$\hat{Q}_n = \frac{1}{n}\sum_{c=1}^{C} \widetilde{\Delta}Z_c'\widetilde{\Delta}Z_c$$
$$\hat{S}_n = \frac{1}{n}\sum_{c=1}^{C} \widetilde{\Delta}Z_c'\widetilde{\Delta}Y_c$$

and thus $\hat{\theta}_{DM} = \hat{Q}_n^{-1}\hat{S}_n$ has the natural interpretation of the least squares estimator of $\widetilde{\Delta}Y_{c\mathbf{i}}$ on $\widetilde{\Delta}Z_{c\mathbf{i}}$. Alternatively, noting that we also have:

$$\hat{Q}_n = \frac{1}{n}\sum_{c=1}^{C} \widetilde{\Delta}Z_c'\Delta Z_c$$
$$\hat{S}_n = \frac{1}{n}\sum_{c=1}^{C} \widetilde{\Delta}Z_c'\Delta Y_c$$
$$\text{where } \Delta\xi_c = (\Delta\xi_{c\mathbf{1}}, \ldots, \Delta Y_{c\mathbf{n_c}})$$

we can view $\hat{\theta}_{DM}$ as an IV estimator at the dyad level that uses $\widetilde{\Delta}Z_{ic}$ as an instrument. To derive consistency, let us make the following additional assumption:

Assumption 15. *Cluster sizes are fixed*

Observe that, Assumption 15 is stronger than the original condition of Hansen and Lee (2019) which permit each cluster to grow in size so long as they remain asymptotically negligible: $\max_{c\leq C} \frac{n_c}{n} \xrightarrow[n\to\infty]{} 0$ (see Assumption 1, p23, Hansen and Lee (2019)). We impose this stronger condition to ensure that the covariates capturing peer influence remain bounded in the *local-aggregate* model. Assumption 15 can be relaxed to match that of Hansen and Lee (2019) in the *linear-in-means* model due to the row-normalization of the adjacency matrix.

Theorem 7. *Suppose Assumptions 7-10, 12 and 15 are satisfied,* $\sup_{c,i}\mathbb{E}(|Y_i^c|^\kappa) < \infty$ *and* $\sup_{c,i}\mathbb{E}(\|Z_i^c\|^\kappa) < \infty$ *for some* $\kappa > 2$. *Then* $\hat{\theta}_{DM} \xrightarrow{p} \theta_0$

Theorem 7 follows directly from Theorem 8 of Hansen and Lee (2019) (see Appendix Section 3.9.8 for a brief discussion). Next, we discuss the asymptotic distribution of $\hat{\theta}$. Define:

$$Q_n = \frac{1}{n}\sum_{c=1}^{C} \mathbb{E}[\widetilde{\Delta}Z_c'\widetilde{\Delta}Z_c]$$
$$\Omega_n = \frac{1}{n}\sum_{c=1}^{C} \mathbb{E}[\widetilde{\Delta}Z_c'\widetilde{\Delta}U_c\widetilde{\Delta}U_c'\widetilde{\Delta}Z_c]$$
$$V_n = Q_n^{-1}\Omega_n Q_n^{-1}$$

Let $\widetilde{\Delta}\hat{U}_c = (\widetilde{\Delta}\hat{U}_{c\mathbf{1}}, \ldots, \widetilde{\Delta}\hat{U}_{c\mathbf{n_c}})$ denote the $n_c \times 1$ vector of residuals where $\widetilde{\Delta}\hat{U}_{c\mathbf{i}} = \mathbb{1}\{D^c_{i_1+} = D^c_{i_2+}\}(\hat{U}^c_{i_1} - \hat{U}^c_{i_2})$ and $\hat{U}^c_{i_k} = Y^c_{i_k} - Z^{c\prime}_{i_k}\hat{\theta}_{DM}$. Define:

$$\hat{\Omega}_n = \frac{1}{n}\sum_{c=1}^{C} \widetilde{\Delta} Z_c' \widetilde{\Delta}\hat{U}_c \widetilde{\Delta}\hat{U}_c' \widetilde{\Delta} Z_c$$

The robust variance estimator is then:

$$\hat{V}_n = \hat{Q}_n^{-1}\hat{\Omega}_n\hat{Q}_n^{-1}$$

The proof of asymptotic normality of our estimator is an adaptation of Theorem 9 in Hansen and Lee (2019) applicable for OLS and 2SLS. It requires stronger conditions than for Theorem 7 regarding the size of the clusters that we collect in Assumption 16 below:

Assumption 16. *Cluster sizes are fixed and for some* $2 \leq \kappa < \infty$

$$\frac{\left(\sum_{c=1}^{C} n_c^{\kappa}\right)^{\frac{2}{\kappa}}}{n} \leq M < \infty$$
$$\max_{r \leq R} \frac{n_c^2}{n} \xrightarrow[n\to\infty]{} 0$$

Theorem 8. *Suppose Assumptions 7-10, 12 are satisfied and 16 holds for some* $2 \leq \kappa < \tau < \infty$. *In addition, suppose that* $\sup_{c,i} \mathbb{E}(|Y_i^c|^{2\tau}) < \infty$ *and* $\sup_{c,i} \mathbb{E}(\|Z_i^c\|^{2\tau}) < \infty$ *and* $\lambda_{min}(\Omega_n) \geq \lambda > 0$ *where* $\lambda_{min}(A)$ *denotes the minimum eigenvalue of* A. *Then:*

$$V_n^{-1/2}\sqrt{n}\left(\hat{\theta}_{DM} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, I_{2K})$$

the robust covariance matrix $\hat{V}_n$ *is consistent in the sense that* $V_n^{-1/2}\hat{V}_n V_n^{-1/2} \xrightarrow{p} I_{2K}$ *so*

$$\hat{V}_n^{-1/2}\sqrt{n}\left(\hat{\theta}_{DM} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, I_{2K})$$

Finally, when all networks are of the same size, $\hat{\theta}_{CF}$ the estimator associated with the observed control function approach is given by:

$$\hat{\theta}_{CF} = \left(\frac{1}{NC}\sum_{c=1}^{C}\sum_{i=1}^{N} S_i^c S_i^{c\prime}\right)^{-1}\left(\frac{1}{NC}\sum_{c=1}^{C}\sum_{i=1}^{N} S_i^c Y_i^{c\prime}\right)$$

where S_i^c is the vector of explanatory variables: $S_i^c = (X_i, \sum_{j\neq i} D_{ij}X_j, D_{i+}, \ldots, D_{i+}^{N-1})'$ for the *local-aggregate* model and $S_i^c = (X_i, \sum_{j\neq i} G_{ij}X_j, D_{i+}, \ldots, D_{i+}^{N-1})'$ for the *linear-in-means*. Since $\hat{\theta}_{CF}$ is a conventional least squares estimator for clustered samples, its asymptotic properties can be directly deduced from Theorem 8 and Theorem 9 of Hansen and Lee (2019). We refer the interested reader to their paper for more details.

3.6.2 SAR and the full model

In SAR (3.12) and the full model (3.7), the *degree-matching* strategy takes the form of a two-stage least squares procedure. To see this, let W_i denote a vector of instruments as in Section 3.4.3. Then,

$$\hat{\theta}_{DM} = \left(\sum_{c=1}^{C} \widetilde{\Delta} Z_c' \widetilde{\Delta} W_c \left(\sum_{c=1}^{C} \widetilde{\Delta} W_c' \widetilde{\Delta} W_c \right)^{-1} \sum_{c=1}^{C} \widetilde{\Delta} W_c' \widetilde{\Delta} Z_c \right)^{-1} \times \left(\sum_{c=1}^{C} \widetilde{\Delta} Z_c' \widetilde{\Delta} W_c \left(\sum_{c=1}^{C} \widetilde{\Delta} W_c' \widetilde{\Delta} W_c \right)^{-1} \sum_{c=1}^{C} \widetilde{\Delta} W_c' \widetilde{\Delta} Y_c \right)$$

The asymptotic behavior of $\hat{\theta}_{DM}$ can once again be derived from Hansen and Lee (2019). We provide a theorem below for completeness but we first require a few additional notations. Now, let:

$$Q_n = \frac{1}{n} \sum_{c=1}^{C} E[\widetilde{\Delta} W_c' \widetilde{\Delta} Z_c], \quad H_n = \frac{1}{n} \sum_{c=1}^{C} E[\widetilde{\Delta} W_c' \widetilde{\Delta} W_c], \quad \Omega_n = \frac{1}{n} \sum_{c=1}^{C} E[\widetilde{\Delta} W_c' \widetilde{\Delta} U_c \widetilde{\Delta} U_c' \widetilde{\Delta} W_c]$$

$$V_n = (Q_n' H_n^{-1} Q_n)^{-1} Q_n' H_n^{-1} \Omega_n H_n^{-1} Q_n (Q_n' H_n^{-1} Q_n)^{-1}$$

$$\hat{Q}_n = \frac{1}{n} \sum_{c=1}^{C} \widetilde{\Delta} W_c' \widetilde{\Delta} Z_c, \quad \hat{H}_n = \frac{1}{n} \sum_{c=1}^{C} \widetilde{\Delta} W_c' \widetilde{\Delta} W_c, \quad \hat{\Omega}_n = \frac{1}{n} \sum_{c=1}^{C} \widetilde{\Delta} W_c' \widetilde{\Delta} U_c \widetilde{\Delta} U_c' \widetilde{\Delta} Z_c$$

$$V_n = (\hat{Q}_n' \hat{H}_n^{-1} \hat{Q}_n)^{-1} \hat{Q}_n' \hat{H}_n^{-1} \hat{\Omega}_n \hat{H}_n^{-1} \hat{Q}_n (\hat{Q}_n' \hat{H}_n^{-1} \hat{Q}_n)^{-1}$$

Theorem 9. *Suppose Assumptions 7-10, 12 are satisfied and 16 holds for some* $2 \leq \kappa < \tau < \infty$. *In addition, suppose that* $\sup_{c,i} \mathbb{E}(|Y_i^c|^{2\tau}) < \infty$, $\sup_{c,i} \mathbb{E}(\|Z_i^c\|^{2\tau}) < \infty$, $\sup_{c,i} \mathbb{E}(\|W_i^c\|^{2\tau}) < \infty$ *and* $\lambda_{min}(\Omega_n) \geq \lambda > 0$ *where* $\lambda_{min}(A)$ *denotes the minimum eigenvalue of* A. *Then:*

$$V_n^{-1/2} \sqrt{n} \left(\hat{\theta}_{DM} - \theta_0 \right) \xrightarrow{d} \mathcal{N}(0, I_{3K})$$

the robust covariance matrix $\hat{V}_n$ *is consistent in the sense that* $V_n^{-1/2} \hat{V}_n V_n^{-1/2} \xrightarrow{p} I_{2K}$ *so*

$$\hat{V}_n^{-1/2} \sqrt{n} \left(\hat{\theta}_{DM} - \theta_0 \right) \xrightarrow{d} \mathcal{N}(0, I_{3K})$$

The proof is omitted as it is identical to that of Theorem 8.

3.7 Monte Carlo Simulations

We examine the properties of our estimators in Monte Carlo Simulations.

Links are generated according to the CID model of Equation (3.2) with graphon $h(x,y) = \Phi\left(\frac{x+y}{\sqrt{2}}\right)$ and with unobserved attributes and dyadic innovations drawn from a standard normal and a standard uniform respectively. This graphon yields an unconditional edge probability of: 50%. For the observed individual covariate, we choose $X_i \sim \text{Bernouilli}(\frac{1}{2})$. Outcomes Y_i are then generated according to one of our baseline models: Equations (3.8) for the *linear-in-means* and Equation (3.9) for the *linear-in-sums* with identical parameters: $\alpha_0 = 0.2, \gamma_0 = 0.5, \delta_0 = 2.0$ and no latent exogenous effects, $\lambda_0 = 0$. This data-generating process is replicated for $C = 200$ independent clusters of size $N = 20, 40, 80$ to produce data that we then use to compute estimates of the social effects. We evaluate the performance of our estimators by repeating this procedure over 5000 Monte Carlo iterations.

Table 3.1: MC Simulation: estimates of $\theta_0 = (\gamma_0, \delta_0)'$ in the baseline *linear-in-means* model

		OLS		PCF		DM		IVJ	
		γ	δ	γ	δ	γ	δ	γ	δ
$N = 20$									
	Bias	-0.000	0.199	0.000	0.029	-0.000	0.002	-0.000	-0.002
	MAB	0.025	0.203	0.012	0.048	0.018	0.068	0.025	0.106
	Size	0.050	0.409	0.052	0.095	0.051	0.056	0.052	0.055
$N = 40$									
	Bias	-0.000	0.099	0.000	0.021	0.000	-0.000	-0.000	-0.000
	MAB	0.018	0.136	0.006	0.046	0.009	0.053	0.018	0.102
	Size	0.051	0.113	0.049	0.071	0.046	0.051	0.051	0.055
$N = 80$									
	Bias	-0.000	0.042	-0.000	0.011	-0.000	-0.000	-0.000	0.003
	MAB	0.013	0.136	0.003	0.046	0.004	0.038	0.013	0.102
	Size	0.054	0.062	0.055	0.054	0.046	0.052	0.054	0.054

NOTES: PCF stands for the proxy control function estimator, DM for the *degree-matching* estimator, IVJ for the instrumental variable strategy proposed in Jochmans (2020). MAB stands for mean absolute bias and Size indicates the fraction of draws that fall outside the asymptotic 95% confidence interval

In Table 3.1, we compare the performance of OLS, a proxy control function estimator adjusting only for degree centrality (i.e we omit the higher order powers), the *degree-matching* estimator and the IV estimator of Jochmans (2020) (henceforth IVJ). The IVJ corresponds to an estimation procedure whereby the average peer characteristics, $\sum_{j\neq i} G_{ij}X_j$ is instrumented by $\sum_{j\neq i}(Q_1)_{ij}X_j$, with:

$$(Q_1)_{ij} = \frac{1}{N-1}\sum_{k\neq i}(G_{-i})_{kj}$$

G_{-i} is the N by N row-normalized adjacency matrix after excluding the links of i

Each matrix entry $(Q_1)_{ij}$ encodes the probability of arriving at j in the network in one step, ruling out any path starting from i. By construction, $(Q_1)_{ij}$ is independent of U_i and since it is correlated with G_{ij} through U_j, the weighted average of peer characteristics $\sum_{j\neq i}(Q_1)_{ij}X_j$ is a valid instrument. We refer the interested reader to Jochmans (2020) for more details on this approach to estimate the social effects.

The left-most columns of Table 3.1 show that the least squares estimates of the exogenous effect suffer from a systematic upward bias, particularly noticeable for $N = 20$, but that recedes with network size. Remarkably, the proxy control function estimator is able to eliminate this bias almost entirely simply by adjusting for degree centrality. The *degree-matching* and IVJ estimates are unsurprisingly the most accurate with nearly no average bias while displaying appropriate size in accordance with theoretical expectations. We note a slight advantage for the *degree-matching* estimator in terms of mean absolute bias in the specific context of this DGP. More generally, it would be interesting to conduct a theoretical comparison of these estimators and determine which one may be more efficient [8]. It is also worth mentioning that estimates of γ_0 are all unbiased, consistent with the fact that X_i is immune to the endogeneity of average peer characteristics by Assumption 8.

In line with our discussion in Section 3.3, Table 3.2 shows that the OLS bias in the *local-aggregate* model is relatively more severe. Indeed, the upward bias of the exogenous effect represents close to 14% of the true parameter value for $N = 20$ and still as much as 5% for $N = 80$. By comparison, the proxy control function and especially the *degree-matching* approach show almost perfect accuracy.

While the simulation results are concordant with the theory for the *degree-matching* estimator, it is interesting to notice that just controlling for degree-centrality in the baseline models does a remarkable job when individuals form links according to a CID process. In practice, this should provide a very easy-to-implement check for applied researchers interested in measuring peer effects when endogenous selection of peers is of potential concern. For networks of moderate size, a significant difference between conventional methods and the proxy-control function approach would be indicative of network endogeneity. To test the latter formally, a Hausman-Durbin-Wu test statistic based on either the *degree-matching* estimator of the control function estimator appears conceivable but I leave the investigation of this question for future work.

[8] Perhaps one advantageous feature of the *degree-matching* procedure within our framework is that it can flexibly accommodate correlated effects and latent exogenous effects while there does not appear to be immediate ways to generalize the approach of Jochmans (2020) to those cases. At the same time, the IVJ estimator is more flexible on other dimensions. For example, it also works for directed networks and could accommodate homophily on observable attributes without requiring the kind of exclusion restrictions imposed in Assumption 14.

Table 3.2: MC Simulation: estimates of $\theta_0 = (\gamma_0, \delta_0)'$ in the baseline *local-aggregate* model

		OLS		PCF		DM	
		γ	δ	γ	δ	γ	δ
$N = 20$							
	Bias	0.001	0.271	0.000	0.000	-0.000	0.000
	MAB	0.021	0.271	0.012	0.006	0.018	0.009
	Size	0.057	1.000	0.052	0.055	0.053	0.049
$N = 40$							
	Bias	0.001	0.175	0.000	-0.000	0.000	0.000
	MAB	0.013	0.175	0.006	0.003	0.009	0.003
	Size	0.062	1.000	0.050	0.051	0.046	0.048
$N = 80$							
	Bias	0.000	0.103	0.000	-0.000	0.000	-0.000
	MAB	0.007	0.103	0.003	0.001	0.004	0.001
	Size	0.056	1.000	0.050	0.057	0.047	0.052

NOTES: PCF stands for the proxy control function estimator, DM for the *degree-matching* estimator. MAB stands for mean absolute bias and Size indicates the fraction of draws that fall outside the asymptotic 95% confidence interval

3.8 Conclusion

In this chapter, we analyze leading peer effect models in the presence of network endogeneity and correlated effects. When the existence of isolated individuals is precluded, identification of the *linear-in-means* model is unaffected by these complications. However, this result crucially hinges on the use of the row-normalized adjacency matrix and is thus not verified in the *local-aggregate* model. Assuming that friendships form according to a CID model, we argue that standard estimation strategies relying on network exogeneity produce biased estimates of the social interaction effects. To address this issue, we introduce two simple methods and derive their asymptotic properties: a control function approach that essentially adjusts for degree centrality in the reduced form outcome equation, and the *degree-matching* approach which takes pairwise differences of agents with identical degree centrality. The common theme of these approaches is that network symmetries can be fruitfully exploited to account for endogenous peer selection and recover the social effects. Finally, results from a Monte Carlo study demonstrate the effectiveness of our estimators and highlight the severe estimation bias that can arise, especially for small networks, when friendship endogeneity is

unaddressed.

3.9 Appendix: proofs and additional materials

3.9.1 SAR(in-means) and the full *linear-in-means* model with no isolated individuals

In this section, we show that under Assumption 11, the spatially autoregressive model (SAR) and the full *linear-in-means* model (Equation (3.4)) are immune to network endogeneity. The SAR model, popular in spatial econometrics posits the following relation:

$$Y_i = \alpha_0 + \beta_0 \sum_{j \neq i} G_{ij} Y_j + \gamma_0 X_i + U_i \tag{3.15}$$

This is a simultaneous-equation model which even in the absence of endogenous peers would pose an endogeneity problem due to the correlation between the average outcome of peers and the unobserved individual attribute, i.e the reflection problem. The standard approach when the covariates and the network are exogenous is to instrument average peer response by the average characteristics of peers. In light of the previous derivations for the baseline *linear-in-means* model, this instrumental variable strategy remains valid under approriate rank conditions if Assumption 11 is verified:

$$Cov(\sum_{j \neq i} G_{ij} X_j, U_i) = 0$$
$$Cov(\sum_{j \neq i} G_{ij} Y_j, \sum_{j \neq i} G_{ij} X_j) \neq 0$$

The full *linear-in-means* model constitutes a more challenging case as we have to deal both with simultaneity and the issue of endogenous peers contaminating the two covariates of peer influence. Rewriting the equations in matrix form, we have:

$$Y = \alpha_0 \iota + \beta_0 GY + \gamma_0 X + \delta_0 GX + U \implies (I - \beta_0 G)Y = \alpha_0 \iota + + \gamma_0 X + \delta_0 GX + U$$

Given our primitive assumption that $|\beta_0| < 1$, and ruling out isolated agents (Assumption 11), we can re-express the system as:

$$Y = \frac{\alpha_0}{(1-\beta_0)} \iota + \gamma_0 X + (\gamma_0 \beta_0 + \delta_0) \sum_{k=0}^{\infty} \beta_0^k G^{k+1} X + \sum_{k=0}^{\infty} \beta_0^k G^k U$$
$$\implies GY = \frac{\alpha_0}{(1-\beta_0)} \iota + \gamma_0 GX + (\gamma_0 \beta_0 + \delta_0) \sum_{k=0}^{\infty} \beta_0^k G^{k+2} X + \sum_{k=0}^{\infty} \beta_0^k G^{k+1} U$$

In a context where the network and the regressors are exogenous, Bramoullé et al. (2009) suggests using G^2X, $G^3X \ldots$ as possible instruments provided that $(\gamma_0 \beta_0 + \delta_0) \neq 0$. It

turns out that these instruments remain valid in our more complicated setting. When, $(\gamma_0\beta_0 + \delta_0) \neq 0$ the reduced-form expression hereinabove shows that GY and the peers peers characteristics G^2X are correlated so relevance of the instrument is satisfied. To confirm the exogeneity of the instrument, observe that since $D_{i+} > 0$ with probability one, we have the following:

$$
\begin{aligned}
&Cov((G^2X)_i, U_i) \\
&= Cov(\sum_{j\neq i} G_{ij} \sum_{k\neq j} G_{jk}X_k, U_i) \\
&= \mathbb{E}\left(\sum_{j\neq i} G_{ij} \sum_{k\neq j} G_{jk}X_kU_i\right) - \mathbb{E}\left(\sum_{j\neq i} G_{ij} \sum_{k\neq j} G_{jk}X_k\right)\mathbb{E}(U_i) \\
&= \left(\mathbb{E}\left(\sum_{j\neq i} G_{ij} \sum_{k\neq j} G_{jk}U_i\right) - \mathbb{E}\left(\sum_{j\neq i} G_{ij} \sum_{k\neq j} G_{jk}\right)\mathbb{E}(U_i)\right)\mathbb{E}(X) \text{ (by Assumptions 7-8)} \\
&= \left(\mathbb{E}\left(U_i \sum_{j\neq i} \frac{D_{ij}}{D_{i+}} \sum_{k\neq j} \frac{D_{jk}}{D_{k+}} \mathbb{1}\{D_{i+} > 0, D_{k+} > 0\}\right)\right. \\
&\left. -\mathbb{E}\left(\sum_{j\neq i} \frac{D_{ij}}{D_{i+}} \sum_{k\neq j} \frac{D_{jk}}{D_{k+}} \mathbb{1}\{D_{i+} > 0, D_{k+} > 0\}\right)\mathbb{E}(U_i)\right) \times \mathbb{E}(X) \\
&= \Big(\mathbb{E}\left(U_i | D_{i+} > 0, D_{k+} > 0\right) - \mathbb{E}(U_i)\Big) P(D_{i+} > 0, D_{k+} > 0)\mathbb{E}(X) \\
&= \left(\mathbb{E}\left(U_i\right) - \mathbb{E}(U_i)\right)\mathbb{E}(X) \\
&= 0
\end{aligned}
$$

Consequently, the usual moment restriction holds:

$$\mathbb{E}\left[(\iota \quad G^2X \quad X \quad GX)'(Y - \alpha_0 - \beta_0 GY - \gamma_0 X - \delta_0 GX\right] = 0$$

It follows that the methodology developed in Bramoullé et al. (2009) is also applicable in a setting with endogenous friendships if the network formation model precludes isolated individuals. Because this assumption is violated in a wide class of link formation models, we suggest novel approaches to estimate social effects in Sections 3.4-3.6 that do not rely on Assumption 11.

3.9.2 Proof of Lemma 15

In this part of the Appendix, we prove Lemma 15. First, by the law of iterated expectations:

$$P(D_1 = d_1) = \int P(D_1 = d_1 | U_1 = u_1) f_{U_1}(u_1) du_1$$

Furthermore, given Assumptions 7 and 8, conditional on U_i: D_{ij} and $D_{ik}, k \neq j$ are independent. Consequently:

$$P(D_1 = d_1) = \int \prod_{j \neq 1} P(D_{1j} = d_{1j} | U_1 = u_1) f_{U_1}(u_1) du_1$$

With the notation $g(u_1) = P(D_{1j} = 1 | U_1 = u_1)$ we can write:
$P(D_{1j} = d_{1j} | U_1 = u_1) = g(u_1)^{d_{1j}} (1 - g(u_1))^{1 - d_{1j}}$. It follows that:

$$\begin{aligned} P(D_1 = d_1) &= \int \left(\prod_{j \neq 1} g(u_1)^{d_{1j}} (1 - g(u_1))^{1 - d_{1j}} \right) f_{U_1}(u_1) du_1 \\ &= \int g(u_1)^{\sum_{j \neq 1} d_{1j}} (1 - g(u_1))^{N - 1 - \sum_{j \neq 1} d_{1j}} f_{U_1}(u_1) du_1 \\ &= \int g(u_1)^{d_{1+}} (1 - g(u_1))^{N - 1 - d_{1+}} f_{U_1}(u_1) du_1 \end{aligned}$$

which makes it clear that $P(D_i = d_i)$ is permutation invariant in its argument as it is a function of the degree of centrality of the agent. In proving the latter, we have also shown that $P(D_1 = d_1 | U_1 = u_1)$ is symmetric in d_1. Therefore, via Bayes rule, the joint density also inherits this property:

$$f_{U_1, D_1}(u_1, d_1) = P(D_1 = d_1 | U_1 = u_1) * f_{U_1}(u_1)$$

which in turn implies

$$f_{U_1 | D_1}(u_1 | d_1) = \frac{f_{U_1, D_1}(u_1, d_1)}{P(D_1 = d_1)}$$

is permutation invariant in $d_1 = (d_{12}, d_{13}, \ldots, d_{1N})$. In other words, the conditional density of U_1 given agent 1' connections in the network: $f_{U_1 | D_{12}, D_{13}, \ldots, D_{1N}}(u_1 | d_{12}, d_{13}, \ldots, d_{1N})$ is a symmetric function of $(d_{12}, d_{13}, \ldots, d_{1N})$.

3.9.3 Proof of Corollary 15.1

Fix any set of links for nodes i and j, $(d_i, d_j) \in \{0,1\}^{N-1} : d_{i+} = d_{j+}$. We can naturally reformulate this network data in terms of a subgraph $S = (V, E, \bar{E})$ with
$V = \{1, \ldots, N\}$, $E = \{(m,k) | m \in \{i,j\}, k \in V(S) : d_{mk} = 1\}$, and
$\bar{E} = \{(m,k) | m \in \{i,j\}, k \in V(S) : d_{mk} = 0\}$. Now, let us partition the set of nodes V as follows: $V = \{i,j\} \cup CF_{ij} \cup EF_i \cup EF_j \cup NF_{ij}$ where

- $CF_{ij} = \{k \in V | d_{ik} = 1 \text{ and } d_{jk} = 1\}$, that is the common friends of i and j
- $EF_i = \{k \in V | d_{ik} = 1 \text{ and } d_{jk} = 0\}$, $EF_j = \{k \in V | d_{jk} = 1 \text{ and } d_{ik} = 0\}$, that is the exclusive friends of i and j respectively. Note that since i and j have the same degree centrality, $|EF_i| = \left|EF_j\right| = m$

- $NF_{ij} = \{k \in V | d_{ik} = 0 \text{ and } d_{jk} = 0\}$, i.e the set of agents that are not linked to i nor j.

It will be convenient to adopt an arbitrary labelling of individuals in EF_i, EF_j as follows: $EF_i = \{i_1, \ldots, i_m\}$, $EF_j = \{j_1, \ldots, j_m\}$ - this is just a technical device. Define $\sigma : V \mapsto V$ by:

1. $\sigma(i) = j$
2. $\forall k \in CF_{ij} \cup NF_{ij}, \sigma(k) = k$
3. $\forall (i_k, j_k) \in EF_i \times EF_j : \sigma(i_k) = j_k$

In words, σ swaps i and j and their exclusive friends and fixes all other agents. By construction, $\sigma \in Aut(S)$ and i,j are automorphic in S. Therefore by the previous corollary

$$U_i | D_i = d_i, D_j = d_j \sim U_j | D_i = d_i, D_j = d_j$$

Likewise $\forall (i_k, j_k) \in EF_i \times EF_j$, i_k, j_k are automorphic in S, thus

$$U_{i_k} | D_i = d_i, D_j = d_j \sim U_{j_k} | D_i = d_i, D_j = d_j$$

Finally, fix $(k, l) \in EF_i \times EF_i$, and consider $\sigma' : V \mapsto V$ defined by:

1. $\sigma'(k) = l$
2. $\forall m \in V \setminus \{k, l\} : \sigma'(m) = m$

By construction σ' is an automorphism and we have shown that k, l are automorphic. Consequently:

$$U_k | D_i = d_i, D_j = d_j \sim U_l | D_i = d_i, D_j = d_j$$

3.9.4 Proof of Theorem 6

We will show that (3.11) holds in two steps. First, observe that in a network of order N, we have:

$$\begin{aligned}
&\mathbb{E}[g(D_i, D_j, X)(\epsilon_i - \epsilon_j)|X, D_{i+} = D_{j+}] \\
&= \mathbb{E}[g(D_i, D_j, X)(\tilde{\epsilon}_i - \tilde{\epsilon}_j)|X, D_{i+} = D_{j+}] \\
&= \frac{\mathbb{E}[g(D_i, D_j, X)(\tilde{\epsilon}_i - \tilde{\epsilon}_j)\mathbb{1}\{D_{i+} = D_{j+}\}|X]}{P(D_{i+} = D_{j+}|X)} \\
&= \frac{\mathbb{E}[g(D_i, D_j, X)(\tilde{\epsilon}_i - \tilde{\epsilon}_j)\mathbb{1}\{D_{i+} = D_{j+}\}|X]}{P(D_{i+} = D_{j+})} \text{ (by Assumptions 7,8,10)} \\
&= \frac{1}{P(D_{i+} = D_{j+})} \\
&\times \mathbb{E}\left[g(D_i, D_j, X)(\tilde{\epsilon}_i - \tilde{\epsilon}_j) \sum_{(d_i,d_j):d_{i+}=d_{j+}} \mathbb{1}\{D_i = d_i, D_j = d_j\}\middle| X\right] \\
&= \sum_{(d_i,d_j):d_{i+}=d_{j+}} \frac{P(D_i = d_i, D_j = d_j)}{P(D_{i+} = D_{j+})} g(d_i, d_j, X) \\
&\times \mathbb{E}[\tilde{\epsilon}_i - \tilde{\epsilon}_j|D_i = d_i, D_j = d_j] \\
&\text{(by Assumptions 7,8,10)}
\end{aligned}$$

This identity is simply a variant of the law of total expectations that we use repeatedly in this chapter. Next, fix $(d_i, d_j) \in \{0,1\}^{N-1} : d_{i+} = d_{j+}$. Appealing to Corollary 15.1, it suffices to show that:

$$\mathbb{E}[\tilde{\epsilon}_i - \tilde{\epsilon}_j|D_i = d_i, D_j = d_j] = 0$$

Case 1: $d_{i+} = d_{j+} = 0$

$$\begin{aligned}
\mathbb{E}[\tilde{\epsilon}_i - \tilde{\epsilon}_j|D_i = d_i, D_j = d_j] &= \mathbb{E}[U_i - U_j|D_i = d_i, D_j = d_j] \\
&= 0 \quad \text{(by Corollary 15.1.1)}
\end{aligned}$$

Case 2: $d_{i+} = d_{j+} > 0$
By definition, conditional on d_i, d_j we have:

$$\begin{aligned}
\tilde{\epsilon}_i - \tilde{\epsilon}_j &= U_i - U_j + \lambda_0 \sum_{k \neq i} \omega_{ik}(d_i) U_k - \lambda_0 \sum_{l \neq j} \omega_{jl}(d_j) U_l \\
&= (1 - \lambda_0 \omega_{ij}(d_i))(U_i - U_j) + \lambda_0 \sum_{k \neq \{i,j\}} \omega_{ik}(d_i) U_k - \lambda_0 \sum_{l \neq \{i,j\}} \omega_{jl}(d_j) U_l \\
&\text{since } (d_{ij} = d_{ji}, d_{i+} = d_{j+} \implies \omega_{ij}(d_i) = \omega_{ji}(d_j)) \\
&= (1 - \lambda_0 \omega_{ij}(d_i))(U_i - U_j) + \lambda_0 \sum_{k \neq \{i,j\}} \omega_{ik}(d_i)(1 - d_{jk}) U_k - \lambda_0 \sum_{l \neq \{i,j\}} \omega_{jl}(1 - d_{il}) U_l
\end{aligned}$$

$$+\lambda_0 \sum_{m\neq\{i,j\}} (\omega_{im}(d_i)D_{jm} - \omega_{jm}(d_j)D_{im})U_k$$
$$= (1-\lambda_0\omega_{ij}(d_i))(U_i - U_j) + \lambda_0 \sum_{k\neq\{i,j\}} \omega_{ik}(d_i)(1-d_{jk})U_k - \lambda_0 \sum_{l\neq\{i,j\}} \omega_{jl}(d_j)(1-d_{il})U_l$$

the last line follows from the fact that the terms involving common friends of i and j cancel out.

Define the sets of "exclusive friends" for i and j respectively:
$EF_i = \{k \in \{1,\ldots,N\} | d_{ik} = 1 \text{ and } d_{jk} = 0\}$, $EF_j = \{k \in \{1,\ldots,N\} | d_{jk} = 1 \text{ and } d_{ik} = 0\}$. Note that since i and j have the same degree centrality, $|EF_i| = \left|EF_j\right| = m$. By Corollary 15.1.1: $\mathbb{E}\left[(1-\lambda_0\omega_{ij}(d_i))(U_i - U_j)|D_i = d_i, D_j = d_j\right] = 0$. Hence,

$$\mathbb{E}[\tilde{\epsilon}_i - \tilde{\epsilon}_j | D_i = d_i, D_j = d_j]$$
$$= \lambda_0 \mathbb{E}\left[\sum_{k\neq\{i,j\}} \omega_{ik}(d_i)(1-d_{jk})U_k - \sum_{l\neq\{i,j\}} \omega_{jl}(d_j)(1-d_{il})U_l \middle| D_i = d_i, D_j = d_j\right]$$

In the *linear-in-means* framework, we have

$$\mathbb{E}[\tilde{\epsilon}_i - \tilde{\epsilon}_j | D_i = d_i, D_j = d_j]$$
$$= \frac{1}{d_{i+}} \sum_{k\in EF_i} \mathbb{E}[U_k | D_i = d_i, D_j = d_j] - \frac{1}{d_{j+}} \sum_{k\in EF_j} \mathbb{E}[U_k | D_i = d_i, D_j = d_j]$$
$$= \frac{1}{d_{i+}} m\mathbb{E}[U_k | D_i = d_i, D_j = d_j; k \in EF_i] - \frac{1}{d_{j+}} m\mathbb{E}[U_k | D_i = d_i, D_j = d_j; k \in EF_j]$$
$$= 0$$

where the penultime line follows from Corollary 15.1.2a, the fact that $d_{i+} = d_{j+}$ and $|EF_i| = \left|EF_j\right| = m$ and the last line is a consequence of Corollary 15.1.2b. The derivations are analogous for the *local-aggregate* version - substitute $\frac{1}{d_{i+}}$ by 1.

3.9.5 Proof of Lemma 16

Without loss of generality, let us focus on the baseline *linear-in-sums* model - the derivations are completely analogous for the baseline *linear-in-means* model. Start with $N = 2$ and let $p = P(D_{ij} = 1)$.

$$\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | D_{i+} = D_{j+}\right]$$
$$= p\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | D_{ij} = 1\right]$$
$$+ (1-p)\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | D_{ij} = 0\right]$$
$$= p\mathbb{E}\left[(X_i - X_j)^2(1,-1)'(1,-1)\right] + (1-p)\mathbb{E}\left[(X_i - X_j)^2(1,0)'(1,0)\right]$$

$$= 2Var(X) \begin{bmatrix} 1 & -p \\ -p & p \end{bmatrix} \text{ (by Assumption 7)}$$

By Assumption 10, the graphon is non-degenerate which rules out $p \in \{0, 1\}$. Thus, the matrix is non singular.
With $N = 3$, let $p = \frac{P(D_{ij}=0, D_{ik}=0, D_{jk}=0)+P(D_{ij}=0, D_{ik}=1, D_{jk}=1)}{P(D_{i+}=D_{j})}$. Then, similar calculations yield:

$$\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | D_{i+} = D_{j+}\right] = 2Var(X) \begin{bmatrix} 1 & -p \\ -p & p \end{bmatrix} \text{ (by Assumption 7)}$$

By Assumption 10, we conclude again that the matrix is non-singular. When $N \geq 4$, there always exist network wirings such that agent i and agent j have the same strictly positive degree and do not share exactly the same set of friends. For instance, in the case of a tetrad, $\mathcal{T} = \{D_{ij} = 0, D_{ik} = 1, D_{il} = 0, D_{jk} = 0, D_{jl} = 1\}$ is such an event and a subset of $D_{i+} = D_{j+}$. Observe now that

$$\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | \mathcal{T}\right] = 2Var(X) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ (by Assumption 7)}$$

so $\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | \mathcal{T}\right]$ is positive definite which immediately implies that $\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | D_{i+} = D_{j+}\right]$ is positive definite as well. This follows from the fact that the latter is a convex combination of positive semi-definite matrices with some of them being strictly positive definite such as $\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | \mathcal{T}\right]$.

3.9.6 Proof of Proposition 6

The method of proof is standard and follows from expanding the squared term in $\mathcal{Q}(\theta)$:

$$\begin{aligned}
\mathcal{Q}(\theta) &= \mathbb{E}\left[\left(\epsilon_i - \epsilon_j - (Z_i - Z_j)'(\theta - \theta_0)\right)^2 | D_{i+} = D_{j+}\right] \\
&= \underbrace{\mathbb{E}\left[\left(\epsilon_i - \epsilon_j\right)^2 | D_{i+} = D_{j+}\right]}_{=T1} + \underbrace{\mathbb{E}\left[(\epsilon_i - \epsilon_j)(Z_i - Z_j)' | D_{i+} = D_{j+}\right]}_{=T2}(\theta - \theta_0) + \\
&\underbrace{(\theta - \theta_0)'\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | D_{i+} = D_{j+}\right](\theta - \theta_0)}_{=T3}
\end{aligned}$$

T_1 is always positive and does not depend on θ.
By Assumption 12: $\mathbb{E}\left[(Z_i - Z_j)(Z_i - Z_j)' | D_{i+} = D_{j+}\right]$ is positive definite, thus T_3 is minimized at $\theta = \theta_0$. Finally,

$$
\begin{aligned}
&\mathbb{E}\left[(\epsilon_i - \epsilon_j)(Z_i - Z_j)|D_{i+} = D_{j+}\right] \\
&= \mathbb{E}\left[\underbrace{\mathbb{E}\left[(Z_i - Z_j)(\epsilon_i - \epsilon_j)|X, D_i, D_j, D_{i+} = D_{j+}\right]}_{=0 \text{ by Theorem 6}}|D_{i+} = D_{j+}\right] \\
&= 0
\end{aligned}
$$

Therefore: $\theta_0 = \arg\min_\theta \mathcal{Q}(\theta)$

3.9.7 Proof of Lemma 17

By symmetry, it suffices to show that

$$\forall m \in \mathbb{N}, \quad \mathbb{E}\left[(G^m X)_i(\epsilon_i - \epsilon_j)\middle| D_{i+} = D_{j+}\right] = \mathbb{E}\left[(G^m X)_i(\tilde{\epsilon}_i - \tilde{\epsilon}_j)\middle| D_{i+} = D_{j+}\right] = 0$$

Recall that $\epsilon_i = A_D + \tilde{\epsilon}_i = A_D + U_i + \sum_{j \neq i} G_{ij} U_j$. The cases $m = 0$ and $m = 1$ follow from Theorem 6 so suppose $m \geq 2$. By the law of total expectations, for a network of order N, we have:

$$
\begin{aligned}
&\mathbb{E}[(G^m X)_i(\tilde{\epsilon}_i - \tilde{\epsilon}_j)|D_{i+} = D_{j+}] \\
&= \sum_{(d_i, d_j): d_{i+} = d_{j+}} \frac{P(D_i = d_i, D_j = d_j)}{P(D_{i+} = D_{j+})} \mathbb{E}\left[(G^m X)_i(\tilde{\epsilon}_i - \tilde{\epsilon}_j)|D_i = d_i, D_j = d_j\right]
\end{aligned}
$$

Thus, it suffices to show that $\forall (d_i, d_j) \in \{0,1\}^{N-1} \times \{0,1\}^{N-1} : d_{i+} = d_{j+}$,

$$\mathbb{E}\left[(G^m X)_i(\tilde{\epsilon}_i - \tilde{\epsilon}_j)|D_i = d_i, D_j = d_j\right] = 0$$

To facilitate the derivations, it is helpful to proceed as in the proof of Theorem 6 and decompose the difference of error terms in two subcomponents:

$$\tilde{\epsilon}_i - \tilde{\epsilon}_j = (1 - \lambda_0 G_{ij})(U_i - U_j) + \lambda_0 \left(\sum_{k \neq \{i,j\}} G_{ik}(1 - D_{jk}) U_k - \sum_{l \neq \{i,j\}} G_{jl}(1 - D_{il}) U_l \right)$$

We will start by showing that $\mathbb{E}\left[(G^m X)_i(U_i - U_j)\middle| D_i = d_i, D_j = d_j\right] = 0$. The mathematical treatment of the second term will be similar.
Fix $(d_i, d_j) \in \{0,1\}^{N-1} \times \{0,1\}^{N-1} : d_{i+} = d_{j+}$ and for notational convenience let $\mathcal{I} = \{D_i = d_i, D_j = d_j\}$. Then:

$$\mathbb{E}\left[(G^m X)_i(U_i - U_j)\middle|\mathcal{I}\right] = \mathbb{E}\left[\sum_{k_1 \neq i} G_{ik_1} \sum_{k_2 \neq k_1} G_{k_1 k_2} \ldots \sum_{k_{m-1} \neq k_m} G_{k_{m-1} k_m} X_{k_m}(U_i - U_j)\middle|\mathcal{I}\right]$$

Case 1: if $d_{i+} = d_{j+} = 0$, then $\mathbb{E}\left[(G^m X)_i (U_i - U_j) \middle| \mathcal{I}\right] = 0$

Case 2: if $d_{i+} = d_{j+} > 0$, then:

$$\mathbb{E}\left[(G^m X)_i (U_i - U_j) \middle| \mathcal{I}\right] = \mathbb{E}(X) \sum_{k_1 : d_{ik_1}=1} \frac{1}{d_{i+}} \underbrace{\mathbb{E}\left[\sum_{k_2 \neq k_1} G_{k_1 k_2} \dots \sum_{k_{m-1} \neq k_m} G_{k_{m-1} k_m} (U_i - U_j) \middle| \mathcal{I}\right]}_{= u_{k_1}}$$

(by Assumptions 7-8)

Abusing notations, define the following sequence of information sets:

$$\begin{aligned}
&\mathcal{I}_{k_1} = \mathcal{I}, \quad \mathcal{F}_{k_1 k_2} = \{D_{k_1 k_2} = 1\} \cup \mathcal{I}_{k_1} \\
&\mathcal{I}_{k_1 k_2} = \sigma\left(D_{k_1(-k_2)}, \mathcal{F}_{k_1 k_2}\right), \quad \mathcal{F}_{k_1 k_2 k_3} = \sigma\left(D_{k_2 k_3} = 1, \mathcal{I}_{k_1 k_2}\right) \\
&\vdots \\
&\mathcal{I}_{k_1, \dots, k_{m-1}} = \sigma\left(D_{k_{m-2}(-k_{m-1})}, \mathcal{F}_{k_1, \dots, k_{m-1}}\right)
\end{aligned}$$

where I use the standard notation $\sigma(W)$ to denote the σ-algebra generated by W. By repeated applications of the law of iterated expectations, we can see that the term u_{k_1} has the following recursive structure:

$$\begin{aligned}
u_{k_1} &= \sum_{k_2 \neq k_1} P(D_{k_1 k_2} = 1 | \mathcal{I}_{k_1}) \mathbb{E}\left[\frac{1}{D_{k_1+}} u_{k_1 k_2} \middle| \mathcal{F}_{k_1 k_2}\right] \\
u_{k_1 k_2} &= \sum_{k_3 \neq k_2} P(D_{k_2 k_3} = 1 | \mathcal{I}_{k_1 k_2}) \mathbb{E}\left[\frac{1}{D_{k_2+}} u_{k_1 k_2 k_3} \middle| \mathcal{F}_{k_1 k_2 k_3}\right] \\
&\vdots \\
u_{k_1, \dots, k_{m-2}} &= \sum_{k_{m-1} \neq k_{m-2}} P(D_{k_{m-2} k_{m-1}} = 1 | \mathcal{I}_{k_1, \dots, k_{m-2}}) \mathbb{E}\left[\frac{1}{D_{k_{m-2}+}} u_{k_1, \dots, k_{m-1}} \middle| \mathcal{F}_{k_1, \dots, k_{m-1}}\right] \\
u_{k_1, \dots, k_{m-1}} &= \mathbb{E}\left[\sum_{k_m \neq k_{m-1}} G_{k_{m-1} k_m} (U_i - U_j) \middle| \mathcal{I}_{k_1, \dots, k_{m-1}}\right]
\end{aligned}$$

Examining the last term of this sequence more closely, we have:

$$u_{k_1, \dots, k_{m-1}} = \mathbb{E}\left[\sum_{k_m \neq k_{m-1}} \frac{D_{k_m}}{D_{k_{m-1}+}} \mathbb{1}\{D_{k_{m-1}+} > 0\} (U_i - U_j) \middle| \mathcal{I}_{k_1, \dots, k_{m-1}}\right]$$

Since we are conditioning on the event $\mathcal{I}_{k_1,\ldots,k_{m-1}}$, we are in particular conditioning on $D_{k_{m-2}k_{m-1}} = 1$ and because, the network is undirected, we know that $D_{k_{m-1}k_{m-2}} = D_{k_{m-2}k_{m-1}} = 1$ which implies $\mathbb{1}\{D_{k_{m-1}+} > 0\} = 1$. Therefore:

$$\begin{aligned} u_{k_1,\ldots,k_{m-1}} &= \mathbb{E}\left[\underbrace{\left(\sum_{k_m \neq k_{m-1}} \frac{D_{k_m}}{D_{k_{m-1}+}}\right)}_{=1}(U_i - U_j)\middle|\mathcal{I}_{k_1,\ldots,k_{m-1}}\right] \\ &= \mathbb{E}\left[(U_i - U_j)\middle|\mathcal{I}_{k_1,\ldots,k_{m-1}}\right] \end{aligned}$$

Going back one step in the sequence, we get:

$$\begin{aligned} & u_{k_1,\ldots,k_{m-2}} \\ &= \sum_{k_{m-1} \neq k_{m-2}} P(D_{k_{m-2}k_{m-1}} = 1|\mathcal{I}_{k_1,\ldots,k_{m-2}})\mathbb{E}\left[\frac{1}{D_{k_{m-2}+}} u_{k_1,\ldots,k_{m-1}}\middle|\mathcal{F}_{k_1,\ldots,k_{m-1}}\right] \\ &= \sum_{k_{m-1} \neq k_{m-2}} P(D_{k_{m-2}k_{m-1}} = 1|\mathcal{I}_{k_1,\ldots,k_{m-2}})\mathbb{E}\left[\frac{1}{D_{k_{m-1}+}}\mathbb{E}\left[(U_i - U_j)\middle|\mathcal{I}_{k_1,\ldots,k_{m-1}}\right]\middle|\mathcal{F}_{k_1,\ldots,k_{m-1}}\right] \\ &= \mathbb{E}\left[\underbrace{\left(\sum_{k_{m-1} \neq k_{m-2}} G_{k_{m-2}k_{m-1}}\right)}_{=1}(U_i - U_j)\middle|\mathcal{I}_{k_1,\ldots,k_{m-2}}\right] \\ &= \mathbb{E}\left[(U_i - U_j)\middle|\mathcal{I}_{k_1,\ldots,k_{m-2}}\right] \end{aligned}$$

and successively

$$\begin{aligned} u_{k_1,\ldots,k_{m-3}} &= \mathbb{E}\left[(U_i - U_j)\middle|\mathcal{I}_{k_1,\ldots,k_{m-3}}\right] \\ &\vdots \\ u_{k_1k_2} &= \mathbb{E}\left[(U_i - U_j)\middle|\mathcal{I}_{k_1,k_2}\right] \\ u_{k_1} &= \mathbb{E}\left[(U_i - U_j)\middle|\mathcal{I}_{k_1}\right] \end{aligned}$$

Thus,

$$\mathbb{E}\left[(G^m X)_i(U_i - U_j)\middle|\mathcal{I}\right] = \mathbb{E}(X)\mathbb{E}\left[(U_i - U_j)\middle|\mathcal{I}\right]$$
$$= 0 \text{ (see Theorem 6)}$$

Now, all that remains to show is:

$$\mathbb{E}\left[(G^m X)_i \left(\sum_{k\neq\{i,j\}} G_{ik}(1 - D_{jk})U_k - \sum_{l\neq\{i,j\}} G_{jl}(1 - D_{il})U_l\right)\middle|\mathcal{I}\right] = 0$$

Case 1: if $d_{i+} = d_{j+} = 0$, this equality is trivially satisfied
Case 2: $d_{i+} = d_{j+} > 0$.
Define the sets $EF_i = \{k \in \{1, \ldots, N\} | d_{ik} = 1 \text{ and } d_{jk} = 0\}$, and
$EF_j = \{k \in \{1, \ldots, N\} | d_{jk} = 1 \text{ and } d_{ik} = 0\}$, i.e the sets of exclusive friends of i and j respectively. Note that since i and j have the same degree centrality, $|EF_i| = \left|EF_j\right| = m$. Then, we equivalently want to show that

$$\frac{1}{d_{i+}} \sum_{k\in EF_i} \mathbb{E}\left[(G^m X)_i U_k\middle|\mathcal{I}\right] - \frac{1}{d_{j+}} \sum_{l\in EF_j} \mathbb{E}\left[(G^m X)_i U_l\middle|\mathcal{I}\right] = 0$$

By repeating exactly the same arguments as above, we get:

$$\mathbb{E}\left[(G^m X)_i U_k\middle|\mathcal{I}; k \in EF_i\right] = \mathbb{E}(X)\mathbb{E}\left[U_k\middle|\mathcal{I}; k \in EF_i\right]$$
$$\mathbb{E}\left[(G^m X)_i U_l\middle|\mathcal{I}; l \in EF_j\right] = \mathbb{E}(X)\mathbb{E}\left[U_l\middle|\mathcal{I}; l \in EF_j\right]$$

Consequently,

$$\frac{1}{d_{i+}} \sum_{k\in EF_i} \mathbb{E}\left[(G^m X)_i U_k\middle|\mathcal{I}\right] - \frac{1}{d_{j+}} \sum_{l\in EF_j} \mathbb{E}\left[(G^m X)_i U_l\middle|\mathcal{I}\right]$$
$$= \frac{\mathbb{E}(X)}{d_{i+}} \left(\sum_{k\in EF_i} \mathbb{E}\left[U_k\middle|\mathcal{I}\right] - \sum_{l\in EF_j} \mathbb{E}\left[U_l\middle|\mathcal{I}\right]\right)$$
$$= \frac{\mathbb{E}(X)}{d_{i+}} m \left(\mathbb{E}[U_k|\mathcal{I}; k \in EF_i] - \mathbb{E}[U_l|\mathcal{I}; l \in EF_j]\right)$$
$$= 0$$

The penultimate line follows from Corollary 15.1.2a and the fact that $d_{i+} = d_{j+}$. The last line is a consequence of Corollary 15.1.2b. Putting these intermediate derivations together, we have $\forall (d_i, d_j) \in \{0,1\}^{N-1} \times \{0,1\}^{N-1} : d_{i+} = d_{j+}$

$$\begin{aligned}
&\mathbb{E}[(G^m X)_i(\tilde{\epsilon}_i - \tilde{\epsilon}_j)|D_i = d_i, D_j = d_j] \\
&= (1 - \lambda_0 g_{ij}) \underbrace{\mathbb{E}\left[U_i - U_j | D_i = d_i, D_j = d_j\right]}_{=0} + \\
&\underbrace{\lambda_0 \, \mathbb{E}\left[\sum_{k \neq \{i,j\}} G_{ik}(1 - D_{jk})U_k - \sum_{l \neq \{i,j\}} G_{jl}(1 - D_{il})U_l \middle| D_i = d_i, D_j = d_j\right]}_{=0} \\
&= 0
\end{aligned}$$

Hence, $\mathbb{E}[(G^m X)_i(\tilde{\epsilon}_i - \tilde{\epsilon}_j)|D_{i+} = D_{j+}] = 0$, which concludes the proof

3.9.8 Proof of Theorem 7

From Assumption 12:

$$\begin{aligned}
Q_n &= \frac{1}{n} \sum_{c=1}^{C} \mathbb{E}[\Delta Z_c' \Delta Z_c] \\
&= \frac{1}{n} \sum_{c=1}^{C} \mathbb{E}\left[\sum_{i=1}^{N_c - 1} \sum_{j=i+1}^{N_c} \mathbb{1}\{D_{i+}^c = D_{j+}^c\}(Z_i^c - Z_j^c)(Z_i^c - Z_j^c)'\right] \\
&= \frac{1}{n} \sum_{c=1}^{C} n_c \mathbb{E}\left[\mathbb{1}\{D_{i+}^c = D_{j+}^c\}(Z_i^c - Z_j^c)(Z_i^c - Z_j^c)'\right]
\end{aligned}$$

is positive definite. Furthermore:

$$\begin{aligned}
\mathbb{E}\left[\left|\widetilde{\Delta} Y_{c\mathbf{i}}\right|^{\kappa}\right] &= \mathbb{E}\left[\mathbb{1}\{\bar{D}_{i_1} = \bar{D}_{i_2}\}\left|Y_{i_1}^c - Y_{i_2}^c\right|^{\kappa}\right] \\
&\leq \mathbb{E}\left[\left|Y_{i_1}^c - Y_{i_2}^c\right|^{\kappa}\right] \\
\mathbb{E}\left[\left\|\widetilde{\Delta} Z_{c\mathbf{i}}\right\|^{\kappa}\right] &= \mathbb{E}\left[\mathbb{1}\{\bar{D}_{i_1} = \bar{D}_{i_2}\}\left\|Z_{i_1}^c - Z_{i_2}^c\right\|\right] \\
&\leq \mathbb{E}[\left\|Z_{i_1}^c - Z_{i_2}^c\right\|^{\kappa}]
\end{aligned}$$

From the triangle inequality and the c_r inequality (convexity), we further have:

$$\mathbb{E}\left[\left|\widetilde{\Delta} Y_{c\mathbf{i}}\right|^{\kappa}\right] \leq 2^{\kappa - 1}\left(\mathbb{E}[|Y_{i_1}|^{\kappa}] + \mathbb{E}[|Y_{i_2}|^{\kappa}]\right)$$

$$\mathbb{E}\left[\left\|\widetilde{\Delta} Z_{c\mathbf{i}}\right\|^{\kappa}\right] \leq 2^{\kappa-1}\left(\mathbb{E}[\|Z_{i_1}\|^{\kappa}] + \mathbb{E}[\|Z_{i_2}\|^{\kappa}]\right)$$

It follows that:

$$\sup_{c,\mathbf{i}} \mathbb{E}\left[\left|\widetilde{\Delta} Y_{c\mathbf{i}}\right|^{\kappa}\right] \leq 2^{\kappa} \sup_{c,i} \mathbb{E}[|Y_i^c|^{\kappa}] < \infty$$
$$\sup_{c,\mathbf{i}} E[\|\Delta Z_{c\mathbf{i}}\|^{\kappa}] \leq 2^{\kappa} \sup_{c,i} \mathbb{E}[\|Z_i^c\|^{\kappa}] < \infty$$

Therefore by Theorem 8 of Hansen and Lee (2019), $\hat{\theta} \xrightarrow{p} \theta_0$.

3.9.9 Proof of Theorem 8

As in the proof of Theorem 7 we have that Q_n is positive definite from Assumption 11 and similarly:

$$\sup_{c,\mathbf{i}} \mathbb{E}[|\Delta Y_{c\mathbf{i}}|^{2\tau}] \leq 2^{2\tau} \sup_{c,i} \mathbb{E}[|Y_i^c|^{2\tau}] < \infty$$
$$\sup_{c,\mathbf{i}} \mathbb{E}[\|\Delta Z_{c\mathbf{i}}\|^{2\tau}] \leq 2^{2\tau} \sup_{c,i} \mathbb{E}[\|Z_i^c\|^{2\tau}] < \infty$$

Therefore by Theorem 9 of Hansen and Lee (2019), we have the desired conclusion.

Bibliography

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

Al-Sadoon, M. M., Li, T., and Pesaran, H. (2017). Exponential class of dynamic binary choice panel data models with fixed effects. *Econometric Reviews*, 36(6-9).

Altonji, J. G. and Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, 73(4):1053–1102.

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.

Aradillas-Lopez, A., Honoré, B. E., and Powell, J. L. (2007). Pairwise difference estimation with nonparametric control variables. *International Economic Review*, 48(4):1119–1158.

Arellano, M. (2003). *Panel data econometrics*. OUP Oxford.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variables estimation of error-component models. *Journal of Econometrics*, 68(1):29 – 51.

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

Arellano, M. and Carrasco, R. (2003b). 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.

Auerbach, E. (2019). Identification and estimation of a partially linear regression model using network data. *arXiv preprint arXiv:1903.09679*.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417.

Bekker, P. and Wansbeek, T. (2001). Identification in parametric models. *A companion to theoretical econometrics*, pages 144–161.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, 87(1):115–143.

Blundell, R. and Bond, S. (2000). Gmm estimation with persistent panel data: an application to production functions. *Econometric Reviews*, 19(3):321 – 340.

Blundell, R., Griffith, R., and Windmeijer, F. (2002). Individual effects and dynamics in count data models. *Journal of econometrics*, 108(1):113–131.

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679.

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

Bonhomme, S., Dano, K., and Graham, B. (2022). Sequential moment restrictions in non-linear panel data models. *Working Paper*.

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.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2020). Peer effects in networks: A survey. *Annual Review of Economics*, 12:603–629.

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

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267.

Card, D. (1996). The effect of unions on the structure of wages: a longitudinal analysis. *Econometrica*, 64(4):957 – 979.

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

Carro, J. M. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics*, 140(2):503–528.

Chamberlain, G. (1979). *Heterogeneity, omitted variable bias, and duration dependence*. Harvard Institute of Economic Research.

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

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

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

Chamberlain, G. (1985c). *Longitudinal Analysis of Labor Market Data*, chapter Heterogeneity, omitted variable bias, and duration dependence, pages 3 – 38. Cambridge University Press, Cambridge.

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

Chamberlain, G. (1992). Comment: sequential moment restrictions in panel data. *Journal of Business and Economic Statistics*, 10(2):20 – 26.

Chamberlain, G. (1993). Feedback in panel data models. *Working Paper*.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica*, 78(1):159–168.

Chamberlain, G. (2022). Feedback in panel data models. *Journal of Econometrics*, 226(1):4 – 20.

Chatterjee, S., Diaconis, P., Sly, A., et al. (2011). Random graphs with a given degree sequence. *Annals of Applied Probability*, 21(4):1400–1435.

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.

Cook, R. J. and Ng, E. T. M. (1997). A logistic-bivariate normal model for overdispersed two-state markov processes. *Biometrics*, 53(1):358 – 364.

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

Cox, D. R. (1958b). The regression analysis of binary sequences. *Journal of the Royal Statistical Society B*, 20(2):215 – 242.

Davezies, L., d'Haultfoeuille, X., and Fougère, D. (2009). Identification of peer effects using group size variation. *The Econometrics Journal*, 12(3):397–413.

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'Haultfoeuille, X., and Mugnier, M. (2020). Fixed effects binary choice models with three or more periods. *arXiv preprint arXiv:2009.08108*.

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.

Dieye, R. and Fortin, B. (2017). Gender peer effects heterogeneity in obesity.

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*.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

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

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

Golsteyn, B. H., Non, A., and Zölitz, U. (2021). The impact of peer personality on academic achievement. *Journal of Political Economy*, 129(4):1052–1099.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica*, 76(3):643–660.

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.

Graham, B. S. (2020). Network data. In *Handbook of Econometrics*, volume 7, pages 111–218. Elsevier.

Graham, B. S., Ridder, G., Thiemann, P. M., and Zamarro, G. (2020). Teacher-to-classroom assignment and student achievement. 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. (2001). The information bound of a dynamic panel logit model with fixed effects. *Econometric Theory*, 17(5):913–932.

Hahn, J. and Kuersteiner, G. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica*, 70(4):1639–1657.

Hansen, B. E. and Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of econometrics*, 210(2):268–290.

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.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

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. and Hu, L. (2004). Estimation of cross sectional and panel data censored regression models with endogeneity. *Journal of Econometrics*, 122(2):293–316.

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. and Lewbel, A. (2002). Semiparametric binary choice panel data models without strictly exogeneous regressors. *Econometrica*, 70(5):2053–2063.

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

Honoré, B. E. and Powell, J. L. (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics*, 64(1-2):241–278.

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.

Jochmans, K. (2020). Peer effects and endogenous social interactions. *arXiv preprint arXiv:2008.07886*.

Johnsson, I. and Moon, H. R. (2015). Estimation of peer effects in endogenous social networks: Control function approach. *Review of Economics and Statistics*, pages 1–51.

Kasahara, H. and Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175.

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.

Lehmann, E. L. and Scheffé, H. (2012). Completeness, similar regions, and unbiased estimation-part i. In *Selected works of EL Lehmann*, pages 233–268. Springer.

Liu, X. and Lee, L.-f. (2010). Gmm estimation of social interaction models with centrality. *Journal of Econometrics*, 159(1):99–115.

Liu, X., Patacchini, E., and Zenou, Y. (2014). Endogenous peer effects: local aggregate or local average? *Journal of Economic Behavior & Organization*, 103:39–59.

Lovász, L. (2012). *Large networks and graph limits*, volume 60. American Mathematical Soc.

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.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

Moffitt, R. A. et al. (2001). Policy interventions, low-level equilibria, and social interactions. *Social dynamics*, 4(45-82):6–17.

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.

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

Olley, S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297.

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

Pigini, C. and Bartolucci, F. (2022). Conditional inference for binary panel data models with predetermined covariates. *Econometrics and Statistics*, 23:83–104.

Powell, J. L. (1994). Estimation of semiparametric models. *Handbook of econometrics*, 4:2443–2521.

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

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pages 577–591.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033.

Selfhout, M., Burk, W., Branje, S., Denissen, J., Van Aken, M., and Meeus, W. (2010). Emerging late adolescent friendship networks and big five personality traits: A social network approach. *Journal of personality*, 78(2):509–538.

Shalizi, C. R. and McFowland III, E. (2016). Estimating causal peer influence in homophilous social networks by inferring latent locations. *arXiv preprint arXiv:1607.06565*.

Ushchev, P. and Zenou, Y. (2020). Social norms in networks. *Journal of Economic Theory*, 185:104969.

Wooldridge, J. M. (1991). Specification testing and quasi-maximum-likelihood estimation. *Journal of Econometrics*, 48(1-2):29–55.

Wooldridge, J. M. (1997). Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory*, 13(5):667–678.