A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator*

Francesco Bartolucci† and Valentina Nigro‡

Abstract

A model for binary panel data is introduced which allows for state dependence and unobserved heterogeneity beyond the effect of available covariates. The model is of quadratic exponential type and its structure closely resembles that of the dynamic logit model. However, it has the advantage of being easily estimable via conditional likelihood with at least two observations (further to an initial observation) and even in the presence of time dummies among the regressors.

Key words: longitudinal data; quadratic exponential distribution; state dependence.

---

*We thank the Co-Editor, Prof. W. K. Newey, and three anonymous Referees for helpful suggestions and insightful comments. We are also grateful to Prof. Franco Peracchi, Università di Roma "Tor Vergata" (IT), and to Prof. Frank Vella, European University Institute of Florence (IT), for their comments and suggestions. Francesco Bartolucci acknowledges the financial support of the "Einaudi Institute for Economics and Finance" - Rome (IT). Most of the article has been developed during the PhD period spent by Valentino Nigro at the University of Rome "Tor Vergata" (program in Econometrics and Empirical Economics) and is part of her PhD thesis.

†Dipartimento di Economia, Finanza e Statistica, Università di Perugia, 06123 Perugia, Italy, *e-mail:* bart@stat.unipg.it

‡Dipartimento di Studi Economico-Finanziari e Metodi Quantitativi, Università di Roma "Tor Vergata", Via Columbia 2, 00133 Roma, Italy, *e-mail:* Valentina.Nigro@uniroma2.it

1 Introduction

Binary panel data are usually analyzed by using a dynamic logit or probit model which includes, among the explanatory variables, the lags of the response variable and has individual-specific intercepts; see Arellano & Honoré (2001) and Hsiao (2005), among others. These models allow us to disentangle the true *state dependence* (i.e. how the experience of an event in the past can influence the occurrence of the same event in the future) from the propensity to experience a certain outcome in all periods, when the latter depends on unobservable factors (see Heckman, 1981a, 1981b). State dependence arises in many economic contexts, such as job decision, investment choice and brand choice and can determine different policy implications. The parameters of main interest of these models are typically those for the covariates and the true state dependence, which are referred to as *structural parameters*. The individual-specific intercepts are referred to as *incidental parameters*; these are of interest only in certain situations, such as when we need to obtain marginal effects and predictions.

In this paper, we introduce a model for binary panel data which closely resembles the dynamic logit model and, as such, allows for state dependence and unobserved heterogeneity between subjects, beyond the effect of the available covariates. The model is a version of the quadratic exponential model (Cox, 1972) with covariates in which: *(i)* the first-order effects depend on the covariates and on an individual-specific parameter for the unobserved heterogeneity; *(ii)* the second-order effects are equal to a common parameter when they are referred to pairs of consecutive response variables and to 0 otherwise. We show that this parameter has the same interpretation that it has in the dynamic logit model in terms of *log-odds ratio*, a measure of association between binary variables which is well known in the statistical literature on categorical data analysis (Agresti, 2002, Ch. 8). For the proposed model we also provide a justification as a latent index model in which the systematic component depends on expectation about future outcomes, beyond the covariates and the lags of the response variable, and the stochastic component has a standard logistic distribution.

An important feature of the proposed model is that, as for the static logit model, the incidental parameters may be eliminated by conditioning on sufficient statistics for these parameters, which correspond to the sums of the response variables at individual level. Using a terminology derived from Rasch (1961), these statistics will be referred to as *total scores*. The resulting conditional likelihood allows us to identify the structural parameters for the covariates and the state dependence with at least two observations (further to an initial observation). The estimator of the structural parameters based on the maximization of this function is $\sqrt{n}$-consistent; moreover, it is simpler to compute than the estimator of Honoré & Kyriazidou (2000) and may be used even in the presence of time dummies. On the basis of a simulation study, whose results are reported in the Supplementary Material file, we also notice that the estimator has good finite sample properties in terms of both bias and efficiency.

The paper is organized as follows. In the next section we briefly review the dynamic logit model for binary panel data. The proposed model is described in Section 3 where we also show that the total scores are sufficient statistics for its incidental parameters. Identification of the structural parameters and the conditional maximum likelihood estimator of these parameters are illustrated in Section 4.

2 Dynamic logit model for binary panel data

In the following, we first review the dynamic logit model for binary panel data and then we discuss conditional inference, and related inferential methods, on its structural parameters.

2.1 Basic assumptions

Let y_{it} be a binary response variable equal to 1 if subject i ($i = 1, \dots, n$) makes a certain choice at time t ($t = 1, \dots, T$) and to 0 otherwise; also let $\boldsymbol{x}_{it}$ be a corresponding vector of strictly exogenous covariates. The standard fixed-effects approach for binary panel data assumes that

$$y_{it} = 1\{y_{it}^* \geqslant 0\}, \quad y_{it}^* = \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \varepsilon_{it}, \qquad i = 1, \dots, n, \quad t = 1, \dots, T, \tag{1}$$

where $1\{\cdot\}$ is the indicator function and y_{it}^* is a latent variable which may be interpreted as utility (or propensity) of the choice. Moreover, the zero-mean random variables ε_{it} represent error terms. Of primary interest are the vector of parameters for the covariates, $\boldsymbol{\beta}$, and the parameter measuring the state dependence effect, γ. These are the structural parameters which are collected in the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma)'$. The individual-specific intercepts α_i are instead the incidental parameters.

The error terms ε_{it} are typically assumed to be independent and identically distributed conditionally on the covariates and the individual-specific parameters and to have a standard logistic distribution. The conditional distribution of y_{it} given α_i, $\boldsymbol{X}_i = (\, \boldsymbol{x}_{i1} \quad \cdots \quad \boldsymbol{x}_{iT} \,)$ and $y_{i0}, \ldots, y_{i,t-1}$ may then be expressed as

$$p(y_{it}|\alpha_i, \boldsymbol{X}_i, y_{i0}, \ldots, y_{i,t-1}) = p(y_{it}|\alpha_i, \boldsymbol{x}_{it}, y_{i,t-1}) = \frac{\exp[y_{it}(\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma)]}{1 + \exp(\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma)}, \tag{2}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. This is a dynamic logit formulation which implies the following conditional distribution of the overall vector of response variables $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$ given $\alpha_i, \boldsymbol{X}_i$ and y_{i0}:

$$p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp(y_{i+}\alpha_i + \sum_t y_{it}\boldsymbol{x}_{it}'\boldsymbol{\beta} + y_{i*}\gamma)}{\prod_t[1 + \exp(\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma)]}, \tag{3}$$

where $y_{i+} = \sum_t y_{it}$ and $y_{i*} = \sum_t y_{i,t-1}y_{it}$, with the sum $\sum_t$ and the product $\prod_t$ ranging over $t = 1, \ldots, T$. The statistic y_{i+} is referred to as the total score of subject i.

For what follows it is important to note that

$$\log \frac{p(y_{it} = 0|\alpha_i, \boldsymbol{X}_i, y_{i,t-1} = 0)p(y_{it} = 1|\alpha_i, \boldsymbol{X}_i, y_{i,t-1} = 1)}{p(y_{it} = 0|\alpha_i, \boldsymbol{X}_i, y_{i,t-1} = 1)p(y_{it} = 1|\alpha_i, \boldsymbol{X}_i, y_{i,t-1} = 0)} = \gamma,$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Thus, the parameter γ for the state dependence corresponds to the conditional log-odds ratio between $(y_{i,t-1}, y_{it})$ for every i and t.

2.2 Conditional inference

As mentioned in Section 1, an effective approach to estimate the model illustrated above is based on the maximization of the conditional likelihood given suitable sufficient statistics.

For the static version of the model, in which the parameter γ is equal to 0, we have that $\boldsymbol{y}_i$ is conditionally independent of α_i given y_{i0}, $\boldsymbol{X}_i$ and the total score y_{i+}, and then $p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i+}) = p(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i+})$. The likelihood based on this conditional probability allows us to identify $\boldsymbol{\beta}$ for $T \geqslant 2$; by maximizing this likelihood we also obtain a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$. Even if referred to a simpler context, this result goes back to Rasch (1961) and was developed by

Andersen (1970). See also Magnac (2004) who characterized other situations in which the total scores are sufficient statistics for the individual-specific intercepts.

Some of the first authors to deal with the conditional approach for the dynamic logit model (γ is unconstrained) were Cox (1958) and Chamberlain (1985). In particular, the latter noticed that when $T = 3$ and the covariates are omitted from the model, $p(\boldsymbol{y}_i|\alpha_i, y_{i0}, y_{i1} + y_{i2} = 1, y_{i3})$ does not depend on α_i for every y_{i0} and y_{i3}. On the basis of this conditional distribution it is therefore possible to construct a likelihood function which depends on the response configurations of only certain subjects (those such that $y_{i1} + y_{i2} = 1$) and which allows us to identify and consistently estimate the parameter γ.

The approach of Chamberlain (1985) was extended by Honoré & Kyriazidou (2000) to the case where, as in (2), the model includes exogenous covariates. In particular, when these covariates are continuous, they proposed to estimate the vector $\boldsymbol{\theta}$ of structural parameters by maximizing a weighted conditional log-likelihood, with weights depending on the individual covariates through a kernel function which must be defined in advance.

Although the weighted conditional approach of Honoré & Kyriazidou (2000) is of great interest, their results about identification and consistency are based on certain assumptions on the support of the covariates which rule out, for instance, time dummies. Moreover, the approach requires careful choice of the kernel function and of its bandwidth, since these choices affect the performance of their estimator. Furthermore, the estimator is consistent as $n \to \infty$, but its rate of convergence to the true parameter value is slower than $\sqrt{n}$, unless in the presence of only discrete covariates. See also Magnac (2004) and Honoré & Tamer (2006).

Even if not strictly related to the conditional approach, it is worth mentioning that a recent line of research investigates dynamic discrete choice models with fixed-effects proposing bias corrected estimators (see Hahn & Newey, 2004; Carro, 2007). Although these estimators are only consistent when the number of time periods goes to infinity, they have a reduced order of the bias without increasing the asymptotic variance. Monte Carlo simulations have shown their good finite sample performance in comparison to the estimator of Honoré & Kyriazidou (2000) even with not very long panels (e.g. seven time periods).

3 Proposed model for binary panel data

In this section, we introduce a quadratic exponential model for binary panel data and we discuss its main features in comparison to the dynamic logit model.

3.1 Basic assumptions

We assume that

$$p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp[y_{i+}\alpha_i + \sum_t y_{it}\boldsymbol{x}_{it}'\boldsymbol{\beta}_1 + y_{iT}(\phi + \boldsymbol{x}_{iT}'\boldsymbol{\beta}_2) + y_{i*}\gamma]}{\sum_{\boldsymbol{z}} \exp[z_+\alpha_i + \sum_t z_t\boldsymbol{x}_{it}'\boldsymbol{\beta}_1 + z_T(\phi + \boldsymbol{x}_{iT}'\boldsymbol{\beta}_2) + z_{i*}\gamma]}, \quad (4)$$

where the sum $\sum_{\boldsymbol{z}}$ ranges over all the possible binary response vectors $\boldsymbol{z} = (z_1, \ldots, z_T)'$; moreover, $z_+ = \sum_t z_t$ and $z_{i*} = y_{i0}z_1 + \sum_{t>1} z_{t-1}z_t$. The denominator, does not depend on $\boldsymbol{y}_i$ and it is simply a *normalizing constant* that we denote by $\mu(\alpha_i, \boldsymbol{X}_i, y_{i0})$. The model may be seen as a version of the quadratic exponential model of Cox (1972) with covariates in which the first-order effect for y_{it} is equal to $\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}_1$ (to which we add $\phi + \boldsymbol{x}_{it}'\boldsymbol{\beta}_2$ when $t = T$) and the second-order effect for (y_{is}, y_{it}) is equal to γ when $t = s + 1$ and to 0 otherwise. The need of a different parametrization of the first-order effect when $t = T$ and $t < T$ will be clarified below.

It is worth noting that the expression for the probability of $\boldsymbol{y}_i$ given in (4) closely resembles that given in (3) which results from the dynamic logit model. From some simple algebra, we also obtain that

$$\log \frac{p(y_{it} = 0|\alpha_i, \boldsymbol{X}_i, y_{i,t-1} = 0)p(y_{it} = 1|\alpha_i, \boldsymbol{X}_i, y_{i,t-1} = 1)}{p(y_{it} = 0|\alpha_i, \boldsymbol{X}_i, y_{i,t-1} = 1)p(y_{it} = 1|\alpha_i, \boldsymbol{X}_i, y_{i,t-1} = 0)} = \gamma$$

for every i and t, and then, under the proposed quadratic exponential model, γ has the same interpretation that it has under the dynamic logit model, i.e. log-odds ratio between each pair of consecutive response variables. Not surprisingly, the dynamic logit model coincides with the proposed model in absence of state dependence ($\gamma = 0$)[1].

The main difference with respect to the dynamic logit is in the resulting conditional distribution of y_{it} given the available covariates $\boldsymbol{X}_i$ and $y_{i0}, \ldots, y_{i,t-1}$. In fact, (4) implies that

$$p(y_{it}|\alpha_i, \boldsymbol{X}_i, y_{i0}, \ldots, y_{i,t-1}) = \frac{\exp\{y_{it}[\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}_1 + y_{i,t-1}\gamma + e_t^*(\alpha_i, \boldsymbol{X}_i)]\}}{1 + \exp[\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}_1 + y_{i,t-1}\gamma + e_t^*(\alpha_i, \boldsymbol{X}_i)]}, \quad (5)$$

[1] It is also possible to show that, up to a correction term, expression (4) is an approximation of that in (3) obtained by a first-order Taylor expansion around $\alpha_i = 0$, $\boldsymbol{\beta} = \boldsymbol{0}$ and $\gamma = 0$.

where, for $t < T$,

$$e_t^*(\alpha_i, \boldsymbol{X}_i) = \log \frac{1 + \exp[\alpha_i + \boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}_1 + e_{t+1}^*(\alpha_i, \boldsymbol{X}_i) + \gamma]}{1 + \exp[\alpha_i + \boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}_1 + e_{t+1}^*(\alpha_i, \boldsymbol{X}_i)]} = \log \frac{p(y_{i,t+1} = 0|\alpha_i, \boldsymbol{X}_i, y_{it} = 0)}{p(y_{i,t+1} = 0|\alpha_i, \boldsymbol{X}_i, y_{it} = 1)}, \quad (6)$$

and,

$$e_T^*(\alpha_i, \boldsymbol{X}_i) = \phi + \boldsymbol{x}'_{iT}\boldsymbol{\beta}_2. \quad (7)$$

Then, for $t = T$ the proposed model is equivalent to a dynamic logit model with a suitable parametrization. The interpretation of this correction term will be discussed in detail in Section 3.2. For the moment it is important to note that the conditional probability depends on present and future covariates, meaning that these covariates are not strictly exogenous (see Wooldridge, 2001, Sec. 15.8.2). The relation between the covariates and the feedback of the response variables vanish when $\gamma = 0$. Consider also that, for $t < T$, the same Taylor expansion mentioned in note 1 leads to $e_t^*(\alpha_i, \boldsymbol{X}_i) \approx 0.5\gamma$. Under this approximation, $p(y_{it}|\alpha_i, \boldsymbol{X}_i, y_{i0}, \ldots, y_{i,t-1})$ does not depend on the future covariates, and then these covariates may be considered strictly exogenous in an approximate sense.

In the simpler case without covariates, the conditional probability of y_{it} becomes:

$$p(y_{it}|\alpha_i, y_{i0}, y_{i1}, \ldots, y_{i,t-1}) = \frac{\exp\{y_{it}[\alpha_i + y_{i,t-1}\gamma + e_t^*(\alpha_i)]\}}{1 + \exp[\alpha_i + y_{i,t-1}\gamma + e_t^*(\alpha_i)]}, \quad t = 1, \ldots, T-1,$$

whereas, for the last period, we have the logistic parametrization

$$p(y_{iT}|\alpha_i, y_{i0}, y_{i1}, \ldots, y_{i,T-1}) = \frac{\exp[y_{iT}(\alpha_i + y_{i,T-1}\gamma)]}{1 + \exp(\alpha_i + y_{i,T-1}\gamma)},$$

where $e_t^*(\alpha_i) = \log \frac{p(y_{i,t+1}=0|\alpha_i, y_{it}=0)}{p(y_{i,t+1}=0|\alpha_i, y_{it}=1)}$, which is zero if and only in absence of state dependence.

Finally, we have to clarify that the possibility to use quadratic exponential models for panel data is already known in the statistical literature; see Diggle *et al.* (2002) and Molenberghs & Verbeke (2004). However, the parametrization adopted in this type of literature, which is different from the one we propose, is sometimes criticized for the lack of a simple interpretation. In contrast, for our parametrization we provide a justification as a latent index model.

3.2 Model justification and related issues

Expression (5) implies that the proposed model is equivalent to the latent index model

$$y_{it} = 1\{y_{it}^* \geqslant 0\}, \quad y_{it}^* = \alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + e_t^*(\alpha_i, \boldsymbol{X}_i) + \varepsilon_{it}, \quad (8)$$

where the error terms ε_{it} are independent and have standard logistic distribution. Assumption (8) is similar to assumption (1) on which the dynamic logit model is based, the main difference being in the correction term $e_t^*(\alpha_i, \boldsymbol{X}_i)$. As is clear from (6), this term may be interpreted as a measure of the effect of the present choice y_{it} on the expected utility (or propensity) at the next occasion $(t+1)$. In the presence of positive state dependence $(\gamma > 0)$, this correction term is positive since making the choice today has a positive impact on the expected utility. Also note that the different definition of $e_t^*(\alpha_i, \boldsymbol{X}_i)$ for $t < T$ and $t = T$ (compare equations (6) and (7)) is motivated by considering that $e_T^*(\alpha_i, \boldsymbol{X}_i)$ has an unspecified form, because it would depend on future covariates not in $\boldsymbol{X}_i$; then, we assume this term to be equal to a linear form of the covariates $\boldsymbol{x}_{iT}$, in a way similar to what suggested by Heckman (1981c) to deal with the initial condition problem.

As suggested by a Referee, it is possible to justify formulation (8), which involves the correction term for the expectation, on the basis of an extension of the job search model described by Hyslop (1999). The latter is based on the maximization of a discounted utility and relies on a budget constraint in which search costs are considered only for subjects who did not participate to the labour market in the previous year. In our extension, subjects who decide to not participate in the current year save an amount of these costs for the next year, but benefit from the amounts previously saved according to the same rule. The reservation wage results modified so that the decision to participate depends on future expectation about the participation state, beyond the past state. This motivates the introduction of the correction term $e_t^*(\alpha_i, \boldsymbol{X}_i)$ in (8), which accounts for the difference between a behavior of a subject having a budget constraint including expectation about future search costs and a subject with a budget constraint not including this expectation.

Two issues that are worth discussing in order to complete the description of the properties of the model are: (*i*) model consistency with respect to marginalizations over a subset of the response variables; (*ii*) how to avoid assumption (7) on the last correction term.

Assume that (4) holds for the T response variables in $\boldsymbol{y}_i$. For the subsequence of responses $\boldsymbol{y}_i^{(T-1)}$, where in general $\boldsymbol{y}_i^{(t)} = (y_{i1}, \ldots, y_{it})'$, we have

$$p(\boldsymbol{y}_i^{(T-1)}|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp[\sum_{t<T} y_{it}(\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}_1) + \sum_{t<T} y_{i,t-1}y_{it}\gamma][1 + \exp(\phi + \boldsymbol{x}_{iT}'\boldsymbol{\delta} + y_{i,T-1}\gamma)]}{\mu(\alpha_i, \boldsymbol{X}_i, y_{i0})},$$

with $\boldsymbol{\delta} = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2$. After some algebra, this expression can be reformulated as

$$p(\boldsymbol{y}_i^{(T-1)}|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp[\sum_{t<T} y_{it}(\alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}_1) + \sum_{t<T} y_{i,t-1}y_{it}\gamma + y_{i,T-1}e_{T-1}(\alpha_i, \boldsymbol{X}_i)]}{\mu_{T-1}(\alpha_i, \boldsymbol{X}_i, y_{i0})}, \quad (9)$$

with

$$e_{T-1}(\alpha_i, \boldsymbol{X}_i) = \frac{1 + \exp(\phi + \boldsymbol{x}_{iT}'\boldsymbol{\delta} + \gamma)}{1 + \exp(\phi + \boldsymbol{x}_{iT}'\boldsymbol{\delta})}$$

and $\mu_{T-1}(\alpha_i, \boldsymbol{X}, y_{i0})$ denoting the normalizing constant, which is equal to the sum of the numerator of (9) for all the possible configurations of the first $T-1$ response variables. Note that $e_{T-1}(\alpha_i, \boldsymbol{X}_i)$ has an interpretation similar to the correction term $e_{T-1}^*(\alpha_i, \boldsymbol{X}_i)$ for the future expectation which is defined above.

When $\gamma = 0$, $e_{T-1}(\alpha_i, \boldsymbol{X}_i) = 0$ and then $p(\boldsymbol{y}_i^{(T-1)}|\alpha_i, \boldsymbol{X}_i, y_{i0}) = p(\boldsymbol{y}_i^{(T-1)}|\alpha_i, \boldsymbol{X}_i^{(T-1)}, y_{i0})$, with $\boldsymbol{X}_i^{(t)} = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{it})'$; the latter probability may be expressed as in (4) and model consistency with respect to marginalization exactly holds. In the other cases, this form of consistency approximately holds, in the sense that substituting $e_{T-1}(\alpha_i, \boldsymbol{X}_i)$ with its linear approximation we obtain a distribution $p(\boldsymbol{y}_i^{(T-1)}|\alpha_i, \boldsymbol{X}_i^{(T-1)}, y_{i0})$ which may be cast into (4). This argument may be iterated to show that, at least approximately, model consistency holds with respect to marginalizations over an arbitrary number of response variables[2]; in this case, the distribution of interest is $p(\boldsymbol{y}_i^{(t)}|\alpha_i, \boldsymbol{X}_i^{(t)}, y_{i0})$ with t smaller than $T-1$.

Finally, assumption (7) on the last correction term $e_T^*(\alpha_i, \boldsymbol{X}_i)$ can be avoided by conditioning the joint distribution on the corresponding outcome y_{iT}. This would remove this correction term since we have

$$p(y_{i1}, \ldots, y_{iT-1}|\alpha_i, \boldsymbol{X}_i, y_{i0}, y_{iT}) = \frac{\exp[\sum_{t<T} y_{it}\alpha_i + \sum_{t<T} y_{it}\boldsymbol{x}_{it}'\boldsymbol{\beta}_1 + \sum_{t<T+1} y_{it}y_{i,t-1}\gamma]}{\sum_{\boldsymbol{z}} \exp[\sum_{t<T} z_t\alpha_i + \sum_{t<T} z_t\boldsymbol{x}_{it}'\boldsymbol{\beta}_1 + \sum_{t<T+1} z_{it}z_{i,t-1}\gamma]}.$$

This conditional version of the proposed model also has the advantage of being consistent across T. However, it would need at least 3 observations (beyond the initial one) to make the model parameters identifiable. Moreover, the conditional estimator results less efficient with respect to the same estimator applied to the initial model.

[2]Simulation results (see the Supplementary Material file) show that, for different values of γ, the bias of the conditional estimator of the structural parameters is negligible and comparable to that resulting from computing these estimators on the complete data sequence.

3.3 Conditional distribution given the total score

The main advantage of the proposed model with respect to the dynamic logit model is that the total scores y_{i+}, $i = 1, \ldots, n$, represent a set of sufficient statistics for the incidental parameters α_i. This is because, for every i, $\boldsymbol{y}_i$ is conditionally independent of α_i given $\boldsymbol{X}_i$, y_{i0} and y_{i+}.

First of all note that, under the assumption (4),

$$\begin{aligned} p(y_{i+}|\alpha_i, \boldsymbol{X}_i, y_{i0}) &= \sum_{\boldsymbol{z}(y_{i+})} p(\boldsymbol{y}_i = \boldsymbol{z}|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \\ &= \frac{\exp(y_{i+}\alpha_i)}{\mu(\alpha_i, \boldsymbol{X}_i, y_{i0})} \sum_{\boldsymbol{z}(y_{i+})} \exp[\sum_t z_t \boldsymbol{x}'_{it}\boldsymbol{\beta}_1 + z_T(\phi + \boldsymbol{x}'_{iT}\boldsymbol{\beta}_2) + z_{i*}\gamma], \end{aligned}$$

where the sum $\sum_{\boldsymbol{z}(y_{i+})}$ is restricted to all the response configurations $\boldsymbol{z}$ such that $z_+ = y_{i+}$. After some algebra, the conditional distribution at issue becomes

$$\begin{aligned} p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0}, y_{i+}) &= \frac{p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0})}{p(y_{i+}|\alpha_i, \boldsymbol{X}_i, y_{i0})} = \\ &= \frac{\exp[\sum_t y_{it}\boldsymbol{x}'_{it}\boldsymbol{\beta}_1 + y_T(\phi + \boldsymbol{x}'_{iT}\boldsymbol{\beta}_2) + y_{i*}\gamma]}{\sum_{\boldsymbol{z}(y_{i+})} \exp[\sum_t z_t\boldsymbol{x}'_{it}\boldsymbol{\beta}_1 + z_T(\phi + \boldsymbol{x}'_{iT}\boldsymbol{\beta}_2) + z_{i*}\gamma]}. \end{aligned} \tag{10}$$

The expression above does not depend on α_i and therefore is also denoted by $p(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+})$. The same happens for the elements of $\boldsymbol{\beta}_1$ corresponding to the covariates which are time constant. To make this clearer, consider that we can divide the numerator and the denominator of (10) by $\exp(y_{i+}\boldsymbol{x}'_{i1}\boldsymbol{\beta}_1)$ and, after rearranging terms, we obtain

$$p(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+}) = \frac{\exp[\sum_{t>1} y_{it}\boldsymbol{d}'_{it}\boldsymbol{\beta}_1 + y_{iT}(\phi + \boldsymbol{x}'_{iT}\boldsymbol{\beta}_2) + y_{i*}\gamma]}{\sum_{\boldsymbol{z}(y_{i+})} \exp[\sum_{t>1} z_t\boldsymbol{d}'_{it}\boldsymbol{\beta}_1 + z_T(\phi + \boldsymbol{x}'_{iT}\boldsymbol{\beta}_2) + z_{i*}\gamma]}, \tag{11}$$

with $\boldsymbol{d}_{it} = \boldsymbol{x}_{it} - \boldsymbol{x}_{i1}$, $t = 2, \ldots, T$. We consequently assume that $\boldsymbol{\beta}_1$ does not include any intercept common to all time occasions and regression parameters for covariates which are time constant; if included, these parameters would not be identified. This is typical of other conditional approaches, such as that of Honoré & Kyriazidou (2000), and of fixed-effects approaches in which the individual intercepts are estimated together with the structural parameters. Similarly, $\boldsymbol{\beta}_2$ must not contain any intercept for the last occasion, since this is already included through ϕ.

4 Conditional inference on the structural parameters

In the following, we introduce a conditional likelihood based on (11). We also provide formal arguments on the identification of the structural parameters via this function and on the asymptotic properties of the estimator resulting from its maximization.

4.1 Structural parameters identification via conditional likelihood

For an observed sample $(\boldsymbol{X}_i, y_{i0}, \boldsymbol{y}_i)$, $i = 1, \dots, n$, the conditional likelihood has logarithm

$$\ell(\boldsymbol{\theta}) = \sum_i 1\{0 < y_{i+} < T\} \log[p_{\boldsymbol{\theta}}(\boldsymbol{y}_i | \boldsymbol{X}_i, y_{i0}, y_{i+})], \tag{12}$$

where the subscript $\boldsymbol{\theta}$ has been added to $p(\cdot|\cdot)$ to remark that this probability, which is defined in (11), depends on $\boldsymbol{\theta}$. Note that in this case $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \phi, \gamma)'$. Also note that the response configurations $\boldsymbol{y}_i$ with sum 0 or T are removed since these do not contain information on $\boldsymbol{\theta}$.

In order to obtain a simple expression for the score and the information matrix corresponding to $\ell(\boldsymbol{\theta})$, consider that (11) may be expressed in the canonical exponential family form as

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}_i | \boldsymbol{X}_i, y_{i0}, y_{i+}) = \frac{\exp[\boldsymbol{u}(y_{i0}, \boldsymbol{y}_i)' \boldsymbol{A}(\boldsymbol{X}_i)' \boldsymbol{\theta}]}{\sum_{\boldsymbol{z}(y_{i+})} \exp[\boldsymbol{u}(y_{i0}, \boldsymbol{z})' \boldsymbol{A}(\boldsymbol{X}_i)' \boldsymbol{\theta}]},$$

where $\boldsymbol{u}(y_{i0}, \boldsymbol{y}_i) = (y_{i2}, \dots, y_{iT}, y_{i*})'$ and

$$\boldsymbol{A}(\boldsymbol{X}_i) = \begin{pmatrix} \boldsymbol{d}_{i2} & \cdots & \boldsymbol{d}_{i,T-1} & \boldsymbol{d}_{iT} & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{x}_{iT} & \boldsymbol{0} \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix},$$

with $\boldsymbol{0}$ denoting a column vector of zeros of suitable dimension. From standard results on exponential family distributions (Barndoff-Nielsen, 1978, Ch. 8), it is easy to obtain

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{\theta}) &= \boldsymbol{\nabla}_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}) = \sum_i 1\{0 < y_{i+} < T\} \boldsymbol{A}(\boldsymbol{X}_i) \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{X}_i, y_{i0}, \boldsymbol{y}_i), \\ \boldsymbol{J}(\boldsymbol{\theta}) &= -\boldsymbol{\nabla}_{\boldsymbol{\theta}\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}) = \sum_i 1\{0 < y_{i+} < T\} \boldsymbol{A}(\boldsymbol{X}_i) \boldsymbol{V}_{\boldsymbol{\theta}}(\boldsymbol{X}_i, y_{i0}, y_{i+}) \boldsymbol{A}(\boldsymbol{X}_i)', \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{X}_i, y_{i0}, \boldsymbol{y}_i) &= \boldsymbol{u}(y_{i0}, \boldsymbol{y}_i) - E_{\boldsymbol{\theta}}[\boldsymbol{u}(y_{i0}, \boldsymbol{y}_i) | \boldsymbol{X}_i, y_{i0}, y_{i+}], \\ \boldsymbol{V}_{\boldsymbol{\theta}}(\boldsymbol{X}_i, y_{i0}, y_{i+}) &= V_{\boldsymbol{\theta}}[\boldsymbol{u}(y_{i0}, \boldsymbol{y}_i) | \boldsymbol{X}_i, y_{i0}, y_{i+}]. \end{aligned}$$

Suppose now that the subjects in the samples are independent of each other with α_i, $\boldsymbol{X}_i$, y_{i0} and $\boldsymbol{y}_i$ drawn, for $i = 1, \dots, n$, from the *true model*

$$f_0(\alpha, \boldsymbol{X}, y_0, \boldsymbol{y}) = f_0(\alpha, \boldsymbol{X}, y_0) p_0(\boldsymbol{y} | \alpha, \boldsymbol{X}, y_0), \tag{13}$$

where $f_0(\alpha, \boldsymbol{X}, y_0)$ denotes the joint distribution of the individual-specific intercept, the covariates $\boldsymbol{X} = (\, \boldsymbol{x}_1 \quad \cdots \quad \boldsymbol{x}_T \,)$ and the initial observation y_0. Furthermore, $p_0(\boldsymbol{y}|\alpha, \boldsymbol{X}, y_0)$ denotes the conditional distribution of the response variables under the quadratic exponential model (4) when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, with $\boldsymbol{\theta}_0$ denoting the true value of its structural parameters. Under this assumption, we have that $\widehat{Q}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta})/n$ converges in probability to $Q_0(\boldsymbol{\theta}) = E_0[\ell(\boldsymbol{\theta})/n] = E_0\{\log[p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{X}, y_0, y_+)]\}$ for any $\boldsymbol{\theta}$, where $E_0(\cdot)$ denotes the expected value under the true model.

By simple algebra, it is possible to show that the first derivative $\boldsymbol{\nabla}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})$ is equal to $\boldsymbol{0}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and that, provided that $E_0[\boldsymbol{A}(X)\boldsymbol{A}(\boldsymbol{X})']$ is of full rank, the second derivative matrix $\boldsymbol{\nabla}_{\boldsymbol{\theta\theta}} Q(\boldsymbol{\theta})$ is always negative definite. This implies that $Q_0(\boldsymbol{\theta})$ is strictly concave with its only maximum at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and, therefore, the vector of structural parameters is identified.

Note that the regularity condition that $E_0[\boldsymbol{A}(\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X})']$ is of full rank, necessary to ensure that $\boldsymbol{\nabla}_{\boldsymbol{\theta\theta}} Q(\boldsymbol{\theta})$ is negative definite, rules out cases of time-constant covariates (see also the discussion in Section 3.3). It is also worth noting that the structural parameters of the model are identified with $T \geqslant 2$, whereas identification of the structural parameters of the dynamic logit model is only possible when $T \geqslant 3$ (Chamberlain, 1993). See also the discussion provided by Honoré & Tamer (2006).

4.2 Conditional maximum likelihood estimator

The conditional maximum likelihood estimator of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', \hat{\phi}, \hat{\gamma})'$, is obtained by maximizing the conditional log-likelihood $\ell(\boldsymbol{\theta})$. This maximum may be found by a simple iterative algorithm, of Newton-Raphson type. At the hth step, this algorithm updates the estimate of $\boldsymbol{\theta}$ at the previous step, $\boldsymbol{\theta}^{(h-1)}$, as $\boldsymbol{\theta}^{(h)} = \boldsymbol{\theta}^{(h-1)} + \boldsymbol{J}(\boldsymbol{\theta}^{(h-1)})^{-1}\boldsymbol{s}(\boldsymbol{\theta}^{(h-1)})$.

Note that the information matrix $\boldsymbol{J}(\boldsymbol{\theta})$ is always non-negative definite since it corresponds to the sum of a series of variance-covariance matrices. Provided $E_0[\boldsymbol{A}(X)\boldsymbol{A}(\boldsymbol{X})']$ is of full rank, $\boldsymbol{J}(\boldsymbol{\theta})$ is also positive definite with probability approaching 1 as $n \to \infty$. Then, we can reasonably expect that $\ell(\boldsymbol{\theta})$ is strictly concave and has its unique maximum at $\hat{\boldsymbol{\theta}}$ in most economic applications, where the sample size is usually large. Since we also have that the parameter space is equal to $\mathbb{R}^k$, with k denoting the dimension of $\boldsymbol{\theta}$, the above algorithm is very simple to implement and usually converges in a few steps to $\hat{\boldsymbol{\theta}}$, regardless of the starting value

$\boldsymbol{\theta}^{(0)}$.

Under the true model (13) and provided that $E_0[\boldsymbol{A}(X)\boldsymbol{A}(\boldsymbol{X})']$ exists and is of full rank, we have that $\hat{\boldsymbol{\theta}}$ exists, is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_0$ and has asymptotic Normal distribution as $n \to \infty$. This results may be proved on the basis of standard asymptotic results (cfr. Th. 2.7 and 3.1 of Newey and McFadden, 1994).

From Newey & McFadden (1994, Sec. 4.2) also derives that the standard errors of the elements of $\hat{\boldsymbol{\theta}}$ may be obtained as the corresponding diagonal elements of $(\hat{\boldsymbol{J}})^{-1}$ under square root. Note that $\hat{\boldsymbol{J}}$ is obtained as a by-product from the Newton-Raphson algorithm described above. These standard errors may be used to construct confidence intervals for the parameters and testing hypotheses on them in the usual way.

In order to study the finite-sample properties of the conditional estimator, we performed a simulation study (for a detailed description see the Supplementary Material file) that closely follows the one performed by Honoré & Kyriazidou (2000). In particular, we first considered a benchmark design under which samples of different size are generated from the quadratic exponential model (4) for 3 and 7 time occasions, only one covariate generated from a Normal distribution and different values of γ between 0.25 and 2. As in Honoré & Kyriazidou (2000), we also considered other scenarios based on more sophisticated designs for the regressors. Under each scenario, we generated a suitable number of samples and for every sample we computed the proposed conditional estimator, whose property were mainly evaluated in terms of *median bias* and *median absolute error* (MAE). We also computed the corresponding standard errors and obtained confidence intervals of different levels for each structural parameter.

On the basis of the simulation study we conclude that, for each structural parameter, the bias of the conditional estimator is always negligible (with the exception of the estimator $\hat{\gamma}$ when n is small); this bias tends to increase with γ and to decrease with n and very quickly with T. Similarly, we observe that the MAE decreases with n at a rate close to $\sqrt{n}$ and much faster with T. This depends on the fact that the number of observations that contribute to the conditional likelihood increases more than proportionally with T, as an increase of T also determines an increase of the *actual sample size*[3]. Moreover, the MAE of the estimator of each parameter

[3] The actual sample size is the number of response configurations $\boldsymbol{y}_i$ such that $0 < y_{i+} < T$. These response configurations contain information on the structural parameters and contribute to $\ell(\boldsymbol{\theta})$; see equation (12).

increases with γ. This is mainly due to the fact that, when γ is positive, its increase implies a decrease of the actual sample size. The simulation results also show that the confidence intervals for the conditional estimator attains the nominal level for each parameter. This confirms the validity of the rule to compute standard errors based on the information matrix $\hat{\boldsymbol{J}}$.

Given the same interpretation of the parameters of the quadratic exponential and the dynamic logit models, it is quite natural to compare the proposed conditional estimator with available estimators of the parameters of the latter model. In particular, the results of our simulation study can be compared with those of Honoré & Kyriazidou (2000). It emerges that our estimator performs better than their estimator in terms of both bias and efficiency. This is mainly due to the fact that the former exploits a larger number of response configurations with respect to the latter. Similarly, our estimator may be compared with the bias corrected estimator proposed by Carro (2007). In this case, we observe that the former performs much better than the latter when the parameter of interest is γ, whereas our estimator performs slightly worse than that of Carro (2007) when the parameters of interest are those in $\boldsymbol{\beta}_1$. However, in taking these conclusions, one must be conscious that the results here compared derive from simulation studies performed under different, although very similar, models.

References

Agresti, A. (2002): *Categorical Data Analysis*. New York: John Wiley & Sons.

Arellano, M. and Honoré, B. (2001): "Panel data models: some recent developments", in *Handbook of Econometrics*, Vol. V, Ed. Heckman J. J. and Leamer E.. Amsterdam: North-Holland.

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

Barndoff-Nielsen, 0. (1978): *Information and Exponential Families in Statistical Theory*. New York: John Wiley & Sons.

Carro, J. (2007): "Estimating dynamic panel data discrete choice models with fixed effects", *Journal of Econometrics*, 140, 503–528.

Chamberlain, G. (1985): "Heterogeneity, omitted variable bias, and duration dependence", in *Longitudinal Analysis of Labor Market Data*, Ed. Heckman J. J. and Singer B.. Cambridge: Cambridge University Press.

Chamberlain, G. (1993): "Feedback in panel data models", unpublished manuscript, Department of Economics, Harvard University.

Cox, D. R. (1958): "The regression analysis of binary sequences", *Journal of the Royal Statistical Society*, 20, 215–242.

Cox, D. R. (1972): "The analysis of multivariate binary data", *Applied Statistics*, 21, 113–120.

Diggle, P. J. and Heagerty, P. and Liang, K.-Y. and Zeger, S. L. (2002): *Analysis of Longitudinal Data*, New York: Oxford University Press.

Hahn, J. and W.K. Newey (2004): "Jackknife and analytical bias reduction for nonlinear panel models", *Econometrica*, 72, 1295–1319.

Heckman, J. J. (1981a): "Statistical models for discrete panel data", in *Structural Analysis of Discrete Data*, Ed. McFadden D. L. and Manski C. A.. Cambridge, MA: MIT Press.

Heckman, J. J. (1981b): "Heterogeneity and state dependence", in *Structural Analysis of Discrete Data*, Ed. McFadden D. L. and Manski C. A.. Cambridge, MA: MIT Press.

Heckman, J. J. (1981c): "The incidental parameter problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process", in *Structural Analysis of Discrete Data*, Ed. McFadden D. L. and Manski C. A.. Cambridge, MA: MIT Press.

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

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

Hsiao, C. (2005): *Analysis of Panel Data, 2nd edition.* New York: Cambridge University Press.

Hyslop, D. R. (1999): "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women", *Econometrica*, 67, 1255–1294.

Magnac, T. (2004): "Panel binary variables and sufficiency: generalizing conditional logit", *Econometrica*, 72, 1859–1876.

Molenberghs, G. and Verbeke, G. (2004): "Meaningful statistical model formulations for repeated measures", *Statistica Sinica*, 14, 989–1020.

Newey W. K. and McFadden D. L. (1994): "Large sample estimation and hypothesis testing", in *Handbook of Econometrics*, Vol. 4, Ed. Engle R. F. and McFadden D. L.. Amsterdam: North-Holland.

Rasch, G. (1961): "On general laws and the meaning of measurement in psychology", *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, 4, 321–333.

Wooldridge, J. M. (2001): *Econometric analysis of cross section and panel data, 1st edition.* The MIT Press.