Dynamic logit model: pseudo conditional likelihood estimation and latent Markov extension

Francesco Bartolucci[1]

Department of Economics, Finance and Statistics

University of Perugia, IT

bart@stat.unipg.it

[1]joint work with Alessio Farcomeni, Univerity of Rome “La Sapienza”, and Valentina Nigro, Bank of Italy, Rome, IT

Outline

- Part I:
  - ▷ Dynamic logit model for binary panel data
  - ▷ Estimation methods: Marginal & Conditional MLE
  - ▷ Dealing with multivariate categorical panel data
  - ▷ Latent Markov model as a tool for model extensions
- Part II:
  - ▷ Pseudo conditional MLE for the univariate model:
    - ∗ Method formulation
    - ∗ Asymptotic properties
    - ∗ Wald test for state dependence
    - ∗ Simulation study & Application
  - ▷ Multivariate extension of the dynamic logit model:
    - ∗ Model assumptions
    - ∗ Maximum likelihood estimation via EM algorithm
    - ∗ An application to labour market data

Dynamic logit model for binary variables

- Basic notation:
  - ▷ n: sample size
  - ▷ T: number of time occasions
  - ▷ y_{it}: binary response variable for subject i at occasion t
  - ▷ $\boldsymbol{x}_{it}$: vector of exogenous covariates for subject i at occasion t

- Basic assumption ($i = 1, \ldots, n$, $t = 1, \ldots, T$):

$$\log \frac{p(y_{it} = 1 | \alpha_i, \boldsymbol{x}_{it})}{p(y_{it} = 0 | \alpha_i, \boldsymbol{x}_{it})} = \alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma$$

  - ▷ α_i: individual-specific parameters for the *unobserved heterogeneity* which may be treated as random or fixed
  - ▷ $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma)$: *structural parameters* which are of greatest interest

Model interpretation

- The model is equivalent to the *latent index model*

$$
\begin{aligned}
y_{it} &= 1\{y_{it}^* > 0\}, \quad i = 1, \dots, n, \quad t = 1, \dots, T, \\
y_{it}^* &= \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \varepsilon_{it},
\end{aligned}
$$

  ▷ ε_{it}: random error term with standard logistic distribution

- γ is of particular interest since it measures the *state dependence* effect, i.e. the effect that experiencing a situation in the present has on the probability of experiencing the same situation in the future (Heckman, 1981) $\Longrightarrow$ important policy implications

- *Spurious state dependence*, i.e. dependence between the responses due to unobservable covariates, is captured by the parameters α_i

Estimation methods

- In a random-effects approach it is natural to follow the (MML) *marginal maximum likelihood* method, based on the marginal prob.

$$p(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}) = \int p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0})dG(\alpha_i|\boldsymbol{X}_i, y_{i0})$$

  - ▷ $\boldsymbol{y}_i = (y_{i1}, ..., y_{iT})$: response vector for subject i
  - ▷ $\boldsymbol{X}_i$: matrix of all covariates in $\boldsymbol{x}_{it}$
  - ▷ $G(\cdot)$: distribution of $\alpha_i|\boldsymbol{X}_i, y_{i0}$

- The MML approach requires to formulate G, typically a normal distribution independent of $\boldsymbol{X}_i$ and y_{i0}; this assumption may be restrictive (Chamberlain, 1982, 1984, Heckman & Singer, 1984)

- It suffers from the *initial condition problem* due to the dependence of y_{i0} on α_i (Heckman, 1981, Wooldridge, 2000, Hsiao, 2005)

- The fixed-effects approach avoids to formulate the distribution of α_i and the initial condition problem. Under this approach we can adopt:
  - ▷ *joint maximum likelihood* (JML) estimation
  - ▷ *conditional maximum likelihood* (CML) estimation

- The JML method consists of maximizing the *joint likelihood* for the parameters α_i and $\boldsymbol{\theta}$ based on the probability

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

  - ▷ $y_{i+} = \sum_t y_{it}, \quad y_{i*} = \sum_t y_{i,t-1}y_{it}$

- The method suffers from the *incidental parameter problem* (Neyman & Scott, 1948); though computational intensive, methods based on corrected score seem promising (Carro, 2007)

- The CML method consists of maximizing the *conditional likelihood* of the structural parameters $\boldsymbol{\theta}$ given a set of sufficient statistics for the incidental parameters α_i

- The method may be applied only in *particular cases*:

  ▷ for the static logit model ($\gamma = 0$) when the total score y_{i+} is a sufficient statistic for α_i (Andersen 1970, 1972)

  ▷ with only discrete covariates having a certain structure (without any constraint on γ); sufficient statistics with a more complex structure need to be used (Charberlain, 1983)

- When the CML method may be applied, it gives rise to a *consistent estimator* of $\boldsymbol{\theta}$ which is usually simple to compute

- The approach was extended to more general cases by Honoré & Kyriazidou (2000)

CML approach of Honoré & Kyriazidou (HK, 2000)

- For $T = 3$, it is based on the maximization of the *weighted conditional log-likelihood*

$$\sum_i 1\{y_{i1} \neq y_{i2}\} K(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i3}) \times \log[p^*(\boldsymbol{y}_i | y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}, \boldsymbol{x}_{i2} = \boldsymbol{x}_{i3})]$$

  - ▷ $p^*(\boldsymbol{y}_i | \boldsymbol{X}_i, y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}, \boldsymbol{x}_{i2} = \boldsymbol{x}_{i3})$: conditional probability of $\boldsymbol{y}_i$ that we would have if $\boldsymbol{x}_{i3}$ was equal to $\boldsymbol{x}_{i2}$
  - ▷ $K(\cdot)$: kernel function for weighting response configurations of the subjects in the sample

- For $T > 3$, the HK-CML estimator of $\boldsymbol{\theta}$ is based on the maximization of a *pairwise* weighted conditional log-likelihood

- Some limits of the HK-CML estimator:
  - ▷ the estimator is *consistent* but it has a slower convergence rate than $\sqrt{n}$
  - ▷ it cannot be applied with *time dummies* or certain types of categorical covariate in the model
  - ▷ the *effective sample size* is much lower than n and this reduces the efficiency (number of subjects who have not degenerate response configuration given the sufficient statistic, i.e. $y_{i1} \neq y_{i2}$)

- ***A pseudo conditional likelihood approach is proposed which is based on approximating the dynamic logit model by a quadratic exponential model (a particular log-linear model) which admits simple sufficient statistics for the incidental parameters***

Multivariate case

- In the *multivariate case*, $r \geq 2$ response variables are observed at each occasion t for each subject i, further to the covariates $\boldsymbol{x}_{it}$

- These data could be analyzed by r *independent dynamic logit models* with specific parameters for each variable, but in this way we would ignore the dependence between variables at the same occasion

- When some response variables have *more than two categories*, the dynamic logit model is not directly applicable and must be extended in order to take into account that these categories may be ordered

- Relevant approaches are those of Ten Have & Morabia (1999), based on a static model for *bivariate binary data*, and that of Todem *et al.* (2007), based on a continuous latent process to model *multivariate ordinal data*

- ***A multivariate extension of the dynamic logit model is proposed which is based on***:
  - ▷ a *multivariate link function* to parametrize the conditional distribution of the vector of response variables given the covariates, the lagged responses and a vector of subject-specific parameters
  - ▷ modeling the vector of subject-specific parameters by a *homogenous Markov chain* to remove the restriction that unobservable covariates have a time-invariant effect

- The resulting model also represents a *generalization of the latent Markov model* of Wiggins (1973) which includes covariates

- The model may be used with categorical response variables with *more than two categories*, possibly ordered, by using suitable logits

Latent Markov (LM) model (Wiggins, 1973)

- This is a model for the analysis of *longitudinal categorical data* which is used in many contexts where the response variables measure a common unobservable characteristic, e.g. psychological and educational measurement, criminology and educational measurement

- Model assumptions:
  - ▷ (*local independence*, LI) for each subject i, the response variables in $\boldsymbol{y}_i$ are conditionally independent given a latent process $\boldsymbol{z}_i = \{z_{it},\, t = 1, \ldots, T\}$
  - ▷ each latent process $\boldsymbol{z}_i$ follows a *first-order homogeneous Markov chain* with state space $\{1, \ldots, k\}$, initial probabilities λ_c and transition probabilities π_{cd}, with $c, d = 1, \ldots, k$

- The literature on LM model is strongly related to that on *hidden Markov* (HM) models (MacDonald & Zucchini, 1997)

- The main difference is that HM models are suitable for *time series* (single long sequence of observations) and the LM model is suitable for panel data (several short sequences of observations)

- Maximum likelihood estimation of the LM model is performed by an *Expectation-Maximization* (EM) algorithm (Dempster *et al.*, 1977) which is based on alternating two steps until convergence:
  - ▷ *E-step*: compute the expected value (given the observed data) of the log-likelihood of the complete data represented by (y_{it}, z_{it}), $i = 1, \ldots, n$, $t = 1, \ldots, T$
  - ▷ *M-step*: maximize the above expected value with respect to the model parameters

- The E-step requires to compute, for each subject, the conditional probability of each latent state at every time occasion (*posterior probabilities*)

- The posterior probabilities can be efficiently computed by a *recursion* taken from the HM literature, which is similar to that used to compute the model likelihood

- ***An extension of this EM algorithm is proposed to estimate the multivariate version of the dynamic logit model***

Comparison with latent class model

LC:

$$\begin{tikzcd}
Y_1 & Y_2 & \cdots & Y_T \\
Z \arrow[u] \arrow[ur] \arrow[urrr] & & &
\end{tikzcd}$$

LM:

$$\begin{tikzcd}
Y_1 & Y_2 & \cdots & Y_T \\
Z_1 \arrow[u] \arrow[r] & Z_2 \arrow[u] \arrow[r] & \cdots \arrow[r] & Z_T \arrow[u]
\end{tikzcd}$$

- The LM model may then be seen as a generalization of the *latent class model* (Lazarsfeld & Henry, 1968) in which the subjects are allowed to move between latent classes

Pseudo condition likelihood estimation

- We propose an estimation method of the structural parameters $\boldsymbol{\theta}$ of the dynamic logit model based on approximating this model by a *quadratic exponential model* (Cox, 1972)

- The parameters of the approximating model have a *similar interpretation* of those of the dynamic logit model (true model)

- Since the approximating model admits simple sufficient statistics for the subject-specific (incidental) parameters, $\boldsymbol{\theta}$ is estimated by maximizing the corresponding conditional likelihood (*pseudo conditional likelihood*)

- *Asymptotic properties* of the estimator are studied by exploiting well-known results on MLE of misspecified models (White, 1981)

The approximating model

- The approximating model is derived from a *linearization* of the log-probability of $\boldsymbol{y}_i$ under the dynamic logit model

$$\begin{aligned}\log[p(\boldsymbol{y}_i|\alpha_i,\boldsymbol{X}_i,y_{i0})] &= y_{i+}\alpha_i+\sum_t y_{it}\boldsymbol{x}'_{it}\boldsymbol{\beta}+y_{i*}\gamma+\\ &- \sum_t \log[1+\exp(\alpha_i+\boldsymbol{x}'_{it}\boldsymbol{\beta}+y_{i,t-1}\gamma)]\end{aligned}$$

- The linearization is based on a *first-order Taylor series expansion* around $\alpha_i=0$, $\boldsymbol{\beta}=\boldsymbol{0}$ and $\gamma=0$ of the non-linear term, obtaining

$$\sum_t \log[1+\exp(\alpha_i+\boldsymbol{x}'_{it}\boldsymbol{\beta}+y_{i,t-1}\gamma)]\approx 0.5\,\tilde{y}_{i+}\gamma+constant,$$

▷ $\tilde{y}_{i+}=\sum_t y_{i,t-1}$

- Probability of $\boldsymbol{y}_i$ under the approximating model:

$$p^*(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp(y_{i+}\alpha_i + \sum_t y_{it}\boldsymbol{x}'_{it}\boldsymbol{\beta} + y_{i*}\gamma - 0.5\tilde{y}_{i+}\gamma)}{\sum_{\boldsymbol{z}} \exp(z_+\alpha_i + \sum_t z_t\boldsymbol{x}'_{it}\boldsymbol{\beta} + z_{i*}\gamma - 0.5\tilde{z}_{i+}\gamma)}$$

  ▷ $\sum_{\boldsymbol{z}}$: sum ranging over all the binary vectors $\boldsymbol{z} = (z_1, \ldots, z_T)$

- Given α_i and $\boldsymbol{X}_i$, the model corresponds to a quadratic exponential model (Cox, 1972) with *second-order interactions* equal to γ, when referred to consecutive response variables, and to 0 otherwise

- The probability of $\boldsymbol{y}_i$ under the approximating model has an expression similar to that under the true model:

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

Interpretation of the approximating model

- Under the approximating model:
  - ▷ for $t = 2, \dots, T$, y_{it} is *conditionally independent* of $y_{i0}, \dots, y_{i,t-2}$, given $\alpha_i, \boldsymbol{X}_i$ and $y_{i,t-1}$ (same property holds under the true model)
  - ▷ for $t = 1, \dots, T$, the *conditional log-odds ratio* for $(y_{i,t-1}, y_{it})$ is given by (same expression holding under the true model)

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

  - ▷ when $t = T$, the *conditional logit* for y_{it} is given by (same expression holding under the true model)

$$\log \frac{p^*(y_{it} = 1|\alpha_i, \boldsymbol{X}_i, y_{i,t-1})}{p^*(y_{it} = 0|\alpha_i, \boldsymbol{X}_i, y_{i,t-1})} = \alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma$$

- for $t = 1, \dots, T-1$, the logit is (similar under the true model)

$$\log \frac{p^*(y_{it} = 1|\alpha_i, \boldsymbol{X}_i, y_{i,t-1})}{p^*(y_{it} = 0|\alpha_i, \boldsymbol{X}_i, y_{i,t-1})} = \alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + e_t(\alpha_i, \boldsymbol{X}_i) - 0.5\gamma$$

- $e_t(\alpha_i, \boldsymbol{X}_i)$: function of α_i and $\boldsymbol{X}_i$ approximately equal to 0.5γ; it is equal to 0 when $\gamma = 0$ (no state dependence)

- The approximating model *coincides with the true model* when $\gamma = 0$

- Under the approximating model, each y_{i+} is a *sufficient statistic* for the incidental parameter α_i $\implies$ the incidental parameters may be removed by conditioning on these statistics

Pseudo CML estimator

- On the basis of the approximating model, we construct a *pseudo conditional log-likelihood*

$$\ell^*(\boldsymbol{\theta}) = \sum_i 1\{0 < y_{i+} < T\}\ell_i^*(\boldsymbol{\theta}), \quad \ell_i^*(\boldsymbol{\theta}) = \log[p^*(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+})]$$

  ▷ $p^*(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+})$: conditional probability of $\boldsymbol{y}_i$ equal to

$$\frac{\exp(\sum_t y_{it}\boldsymbol{x}'_{it}\boldsymbol{\beta} - 0.5\tilde{y}_{i+}\gamma + y_{i*}\gamma)}{\sum_{\boldsymbol{z}:z_+=y_{i+}} \exp(\sum_t z_t\boldsymbol{x}'_{it}\boldsymbol{\beta} - \sum_t 0.5z_{i,t-1}\gamma + z_{i*}\gamma)}$$

  ▷ $\sum_{\boldsymbol{z}:z_+=y_{i+}}$: sum ranging over all the binary vectors $\boldsymbol{z} = (z_1, \ldots, z_T)$ with same total as $\boldsymbol{y}_i$

- Maximization of $\ell^*(\boldsymbol{\theta})$ is possible by a simple NR algorithm, resulting in the *pseudo* CML *estimator* $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\gamma})$ of the structural parameters

Improved pseudo CML estimator

- This relies on a sharper approximation of the true model based on a *first-order Taylor series expansion* around $\alpha_i = 0$, $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}$ and $\gamma = 0$:

$$\sum_t \log[1 + \exp(\alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma)] \approx \sum_t q_{it} y_{i,t-1}\gamma + constant,$$

  - ▷ $\bar{\boldsymbol{\beta}}$: fixed value of $\boldsymbol{\beta}$ chosen by a preliminary estimation of this parameter vector
  - ▷ $q_{it} = \dfrac{\exp(\boldsymbol{x}'_{it}\bar{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}'_{it}\bar{\boldsymbol{\beta}})}$; it is equal to 0.5 when $\bar{\boldsymbol{\beta}} = \mathbf{0}$

- Probability of $\boldsymbol{y}_i$ under the improved approximating model:

$$p^\dagger(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp(y_{i+}\alpha_i + \sum_t y_{it}\boldsymbol{x}'_{it}\boldsymbol{\beta} - \sum_t q_{it}y_{i,t-1}\gamma + y_{i*}\gamma)}{\sum_{\boldsymbol{z}} \exp(z_+\alpha_i + \sum_t z_t\boldsymbol{x}'_{it}\boldsymbol{\beta} - \sum_t q_{it}z_{i,t-1}\gamma + z_{i*}\gamma)}$$

- The improved approximating model has properties very similar to the approximating model for what concerns its *interpretation* in connection with the true model

- y_{i+} is still a *sufficient statistic* for α_i; the incidental parameters α_i may be removed by conditioning on these sufficient statistics

- An *improved pseudo conditional log-likelihood* results:

$$\ell^\dagger(\boldsymbol{\theta}) = \sum_i 1\{0 < y_{i+} < T\}\ell_i^\dagger(\boldsymbol{\theta}), \quad \ell_i^\dagger(\boldsymbol{\theta}) = \log[p^\dagger(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+})]$$

  ▷ $p^\dagger(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+})$: conditional probability of $\boldsymbol{y}_i$ given y_{i+}

- Maximization of $\ell^\dagger(\boldsymbol{\theta})$ is performed via NR, once $\bar{\boldsymbol{\beta}}$ has been fixed at $\hat{\boldsymbol{\beta}}$, obtaining the *improved pseudo* CML *estimator* $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\gamma})$ of the structural parameters (this substitution may be iterated)

Asymptotic properties (basic pseudo CML estimator)

- We assume an *$i.i.d.$ sampling scheme* from

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

  - ▷ $p_0(\boldsymbol{y}|\alpha, \boldsymbol{X}, y_0)$: conditional distribution of the response variables under the true model with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$
  - ▷ $f_0(\alpha, \boldsymbol{X}, y_0)$: true distribution of $(\alpha, \boldsymbol{X}, y_0)$

- Following Akaike (1973) and White (1981), we define the *pseudo true parameter* vector $\boldsymbol{\theta}_* = (\boldsymbol{\beta}_*, \gamma_*)$ as the $\boldsymbol{\theta}$ which minimizes the KL distance between the true and the approximating models:

$$K^*(\boldsymbol{\theta}) = E_0\{\log[p_0(\boldsymbol{y}|\alpha, \boldsymbol{X}, y_0)/p^*_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{X}, y_0, y_+)]\}$$

Theorem 1. *For $T \geq 2$ and provided that $E_0(\boldsymbol{X}\boldsymbol{D}'\boldsymbol{D}\boldsymbol{X}')$ exists and is of full rank, with $\boldsymbol{D} = (\,-\mathbf{1} \quad \boldsymbol{I}\,)$, as $n \to \infty$ we have:*

- *(Existence) $\hat{\boldsymbol{\theta}}$ exists with probability approaching 1*
- *(Consistency) $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_*$*
- *(Normality) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{V}_0(\boldsymbol{\theta}_*))$*
  - ▷ $\boldsymbol{V}_0(\boldsymbol{\theta}) = \boldsymbol{J}_0(\boldsymbol{\theta})^{-1}\boldsymbol{S}_0(\boldsymbol{\theta})\boldsymbol{J}_0(\boldsymbol{\theta})^{-1}$
  - ▷ $\boldsymbol{J}_0(\boldsymbol{\theta}) = E_0[\boldsymbol{\nabla}_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell_i^*(\boldsymbol{\theta})]$
  - ▷ $\boldsymbol{S}_0(\boldsymbol{\theta}) = E_0[\boldsymbol{\nabla}_{\boldsymbol{\theta}}\ell_i^*(\boldsymbol{\theta})\boldsymbol{\nabla}_{\boldsymbol{\theta}}\ell_i^*(\boldsymbol{\theta})']$
- *(Sandwich variance estimation) $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \boldsymbol{V}_0(\boldsymbol{\theta}_*) \implies$ we can compute consistent $s.e.(\hat{\boldsymbol{\theta}})$*

- The pseudo true parameter $\boldsymbol{\theta}_*$ is equal to $\boldsymbol{\theta}_0$ when $\gamma_0 = 0$ (*no state dependence*) since in this case the approximating model coincides with the true model $\implies \hat{\boldsymbol{\theta}}$ is consistent

- In the *other cases* ($\gamma_0 \neq 0$), we expect $\boldsymbol{\theta}_*$ to be reasonably close $\boldsymbol{\theta}_0$ $\implies$ the pseudo CML estimator is "quasi consistent"

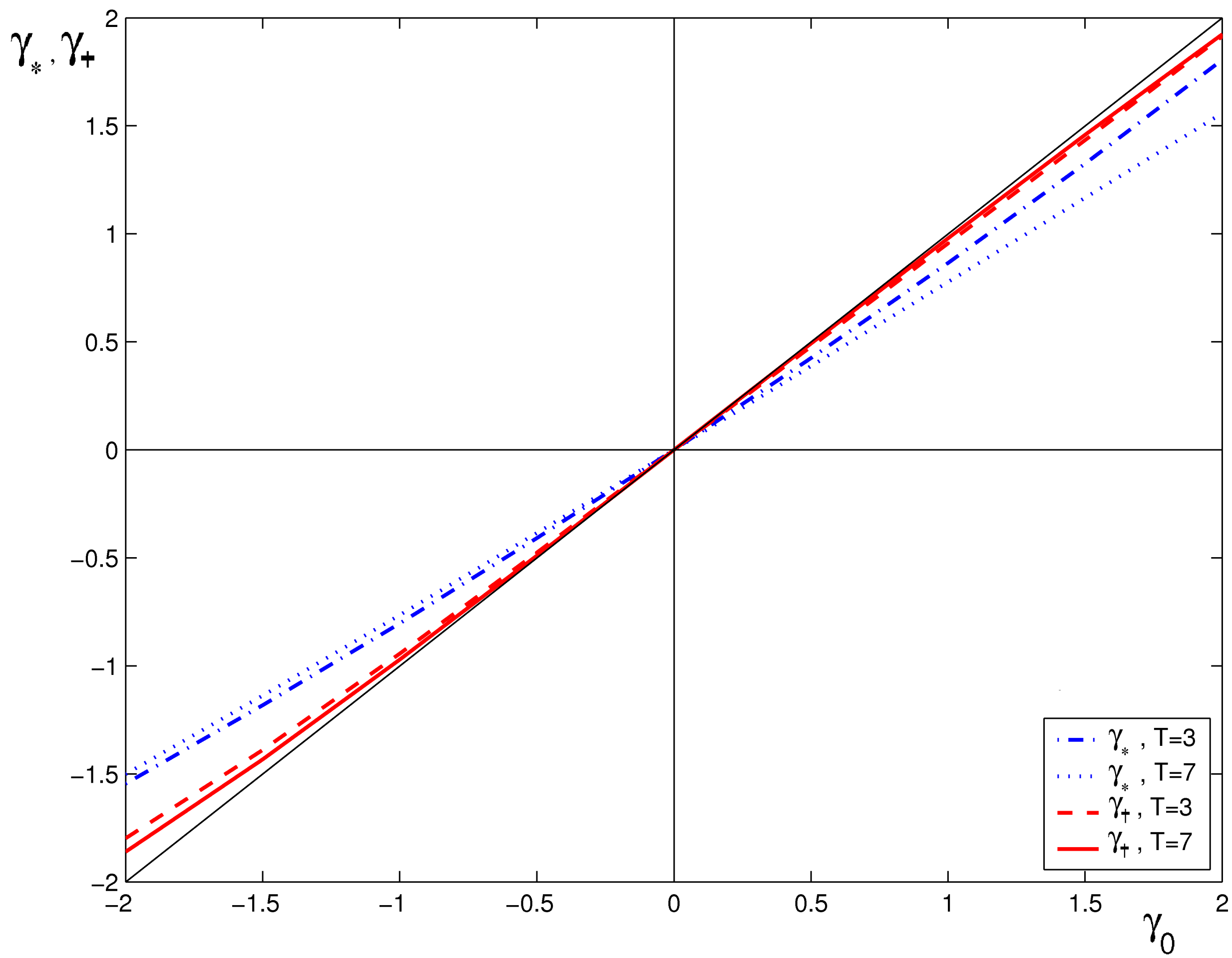
- Similar properties hold for the *improved pseudo CML estimator*, which converges to the *pseudo true parameter* vector $\boldsymbol{\theta}_\dagger = (\boldsymbol{\beta}_\dagger, \gamma_\dagger)$ corresponding to the minimum of the KL distance

$$K^\dagger(\boldsymbol{\theta}) = E_0\{\log[p_0(\boldsymbol{y}|\alpha, \boldsymbol{X}, y_0)/p^\dagger_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{X}, y_0, y_+)]\}$$

- We expect the $\boldsymbol{\theta}_\dagger$ to be closer to $\boldsymbol{\theta}_0$ with respect to $\boldsymbol{\theta}_*$; this is graphically illustrated for certain particular cases

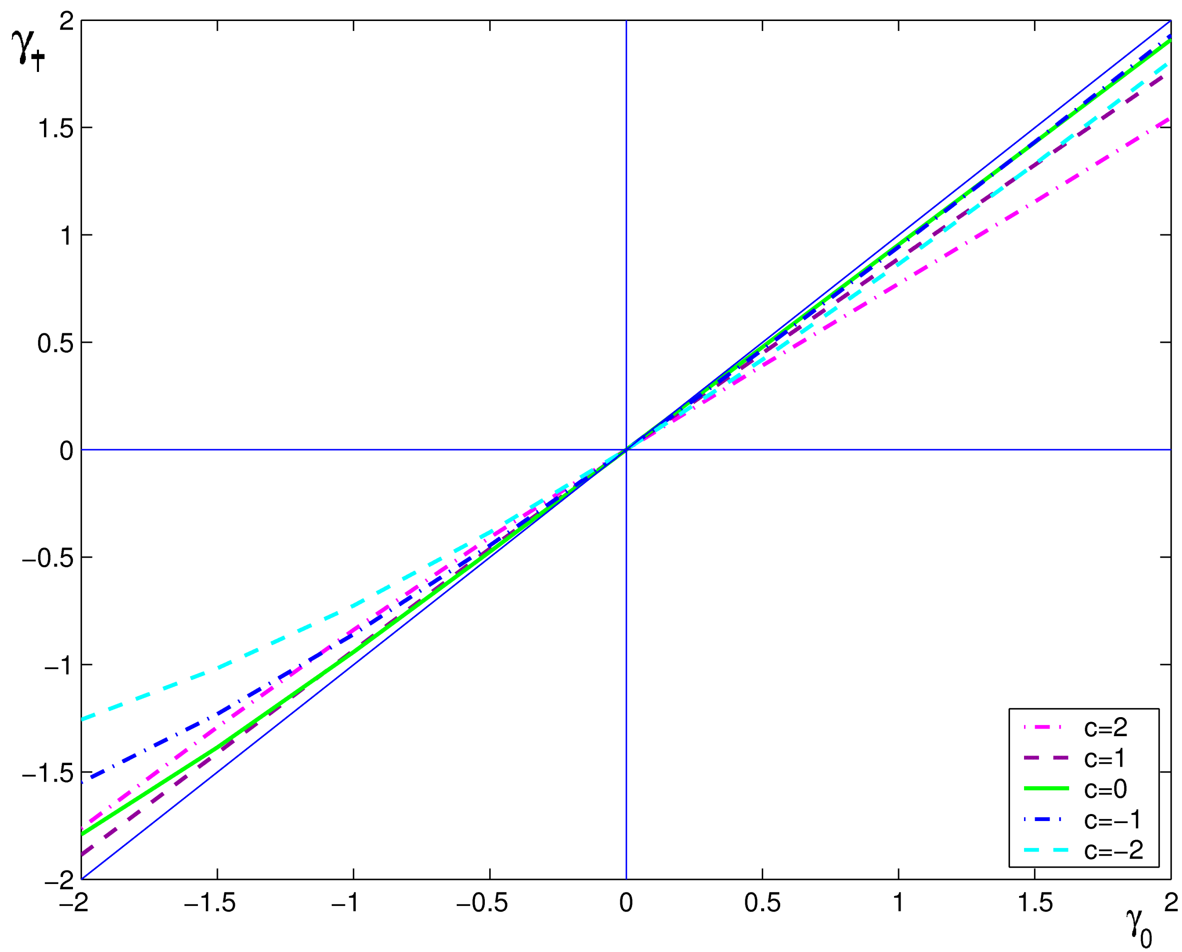
Graphical illustration (1/3)

- Values of the pseudo true parameters γ_* and $\gamma_\dagger$ for different values of the true parameter for state dependence γ_0 and different time periods, with $\beta = 1$, $x_{it} \sim N(0, \pi^2/3)$, $\alpha_i = (x_{i0} + \sum_t x_{it})/(T+1)$



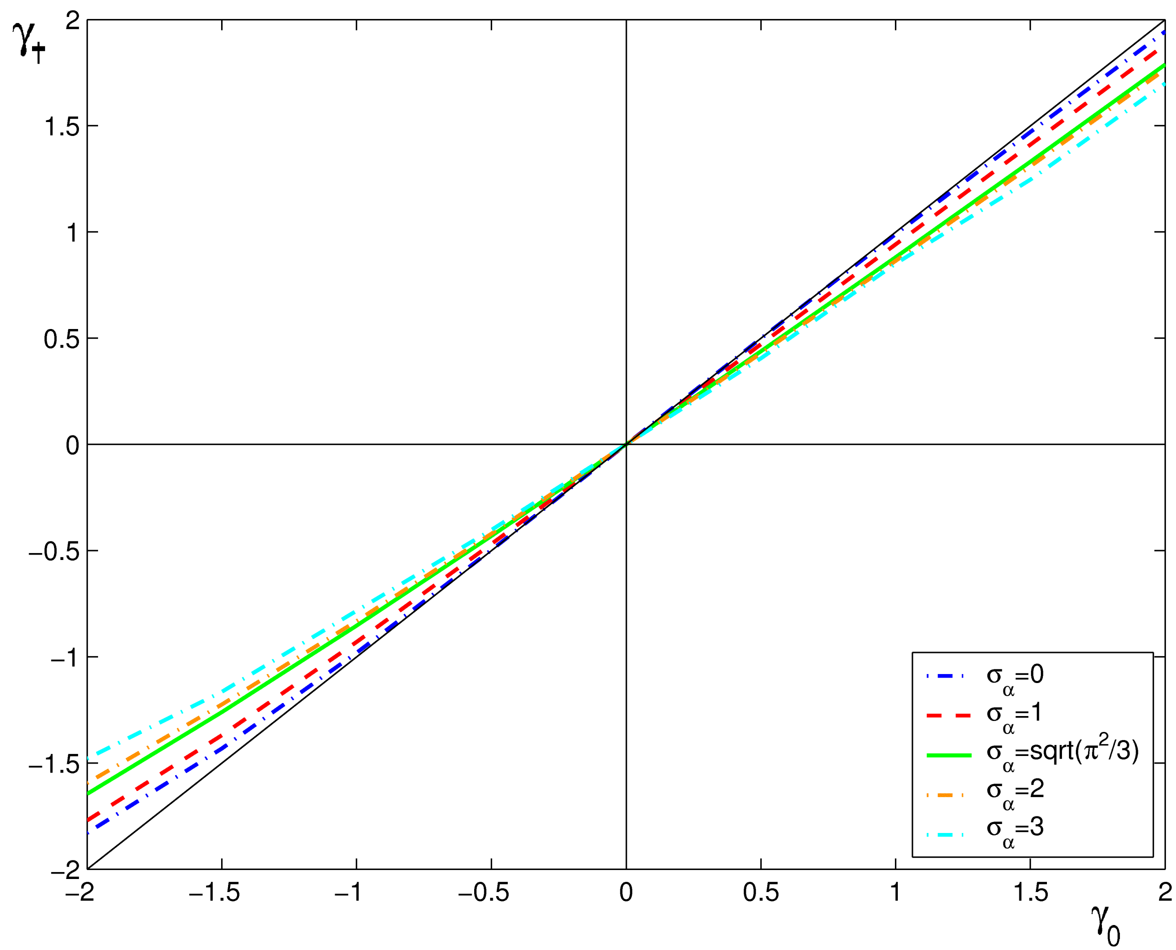
Graphical illustration (2/3)

- Values of the pseudo true parameter $\gamma_\dagger$ for different values of the true parameter for the state dependence γ_0, with $T=3$, $\beta=1$, $x_{it} \sim N(0, \pi^2/3)$, $\alpha_i = c + (x_{i0} + \sum_t x_{it})/(T+1)$



Graphical illustration (3/3)

- Values of the pseudo true parameter $\gamma_\dagger$ for different values of the true parameter for state dependence γ_0, with $T = 3$, $\beta = 1$, $x_{it} \sim N(0, \pi^2/3)$, $\alpha_i \sim N(0, \sigma^2_\alpha)$



Simulation study

- *Finite-sample properties* are studied by simulation under different settings (model for the covariates, sample size n, number of time occasions T, true value of the parameters of the dynamic logit model)

- The basic and improved pseudo CML estimators have *negligible bias* and *same efficiency* when γ_0 is close to 0

- When γ_0 is significantly different from 0, the improved estimator is *considerably more efficient* than the basic estimator and has a much lower bias

- *Confidence intervals* based on the improved estimator usually attain the nominal coverage level even for γ_0 very far from 0

Comparison with the HK estimator

- Simulation based on 1000 samples drawn from the dynamic logit model with $x_{it} \sim N(0, \pi^2/3)$, $\alpha_i = (x_{i0} + \sum_t x_{it})/(T+1)$, $\beta = 1$, $\gamma = 0.5, 2$ for different values of n and T

- Comparison in terms of *median bias* (Bias) and *median absolute error* (MAE)

γ	T	n	Estimator	Parameter β		Parameter γ	
				Bias	MAE	Bias	MAE
0.5	3	250	Weighted HK	0.076	0.154	-0.039	0.403
(37% - 57%)			*Improved pseudo*	*0.010*	*0.086*	*-0.027*	*0.239*
		1000	Weighted HK	0.038	0.086	-0.035	0.178
			Improved pseudo	*0.002*	*0.045*	*-0.017*	*0.125*
		4000	Weighted HK	0.019	0.044	-0.035	0.102
			Improved pseudo	*0.000*	*0.023*	*-0.021*	*0.066*

				Parameter β		Parameter γ	
γ	T	n	Estimator	Bias	MAE	Bias	MAE
0.5	7	250	Weighted HK	0.014	0.050	-0.053	0.131
(43% - 91%)			*Improved pseudo*	*0.001*	*0.039*	*-0.009*	*0.107*
		1000	Weighted HK	0.009	0.027	-0.041	0.075
			Improved pseudo	*-0.001*	*0.021*	*-0.013*	*0.058*
		4000	Weighted HK	0.005	0.015	-0.033	0.039
			Improved pseudo	*0.001*	*0.010*	*-0.010*	*0.027*
2	3	250	Weighted HK	0.196	0.251	-0.056	0.620
(26% - 42%)			*Improved pseudo*	*0.015*	*0.111*	*-0.056*	*0.369*
		1000	Weighted HK	0.113	0.136	-0.148	0.321
			Improved pseudo	*-0.008*	*0.051*	*-0.083*	*0.166*
		4000	Weighted HK	0.063	0.074	-0.118	0.163
			Improved pseudo	*-0.006*	*0.027*	*-0.079*	*0.104*
2	7	250	Weighted HK	0.016	0.064	-0.195	0.227
(34% - 76%)			*Improved pseudo*	*0.001*	*0.046*	*-0.072*	*0.133*
		1000	Weighted HK	0.016	0.034	-0.160	0.164
			Improved pseudo	*-0.002*	*0.024*	*-0.066*	*0.083*
		4000	Weighted HK	0.006	0.017	-0.116	0.116
			Improved pseudo	*-0.001*	*0.012*	*-0.066*	*0.067*

- The improved pseudo CML estimator *outperforms* the HK estimator (this is evident even in other settings)

- The *advantage in terms of efficiency* is greater for shorter panels ($T = 3$ instead of $T = 7$) and for higher values of γ ($\gamma = 2$ instead of $\gamma = 0.5$), but is rather insensitive to n

- The advantage can be explained considering that the *actual sample size* is much higher under the proposed approach than in the HK approach

- The proposed estimator is also much *simpler to compute* than the HK estimator and can be used with $T \geq 2$ instead of $T \geq 3$ and with no limitations on the covariate structure

Wald test for state dependence

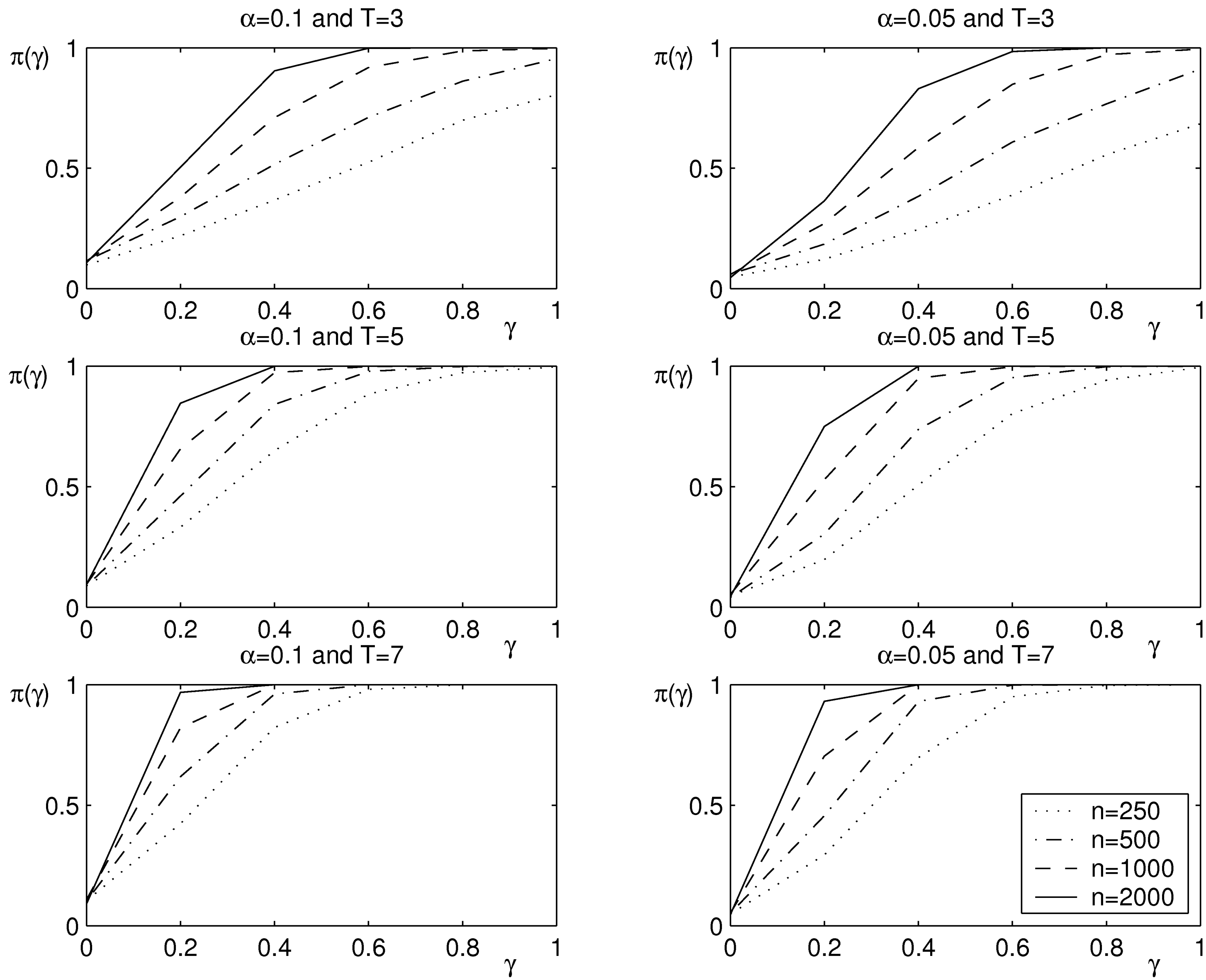
- Since the basic pseudo CML estimator is consistent when $\gamma_0 = 0$, we can exploit this estimator to construct a Wald test for the hypothesis of *absence of state dependence* $(H_0 : \gamma = 0)$ based on the statistic:

$$t = \frac{\hat{\gamma}}{s.e.(\hat{\gamma})}$$

- The power of this test was studied by a *simulation* in which samples were drawn from the dynamic logit model with $x_{it} \sim N(0, \pi^2/3)$, $\alpha_i = (x_{i0} + \sum_t x_{it})/(T+1)$, $\beta = 1$, γ between 0 and 1

- The results show that the nominal *nominal significant level* (α) is attained under the null hypothesis and that the power has a typical behavior (increases with n, α and the distance of the true γ from 0)

Simulation results for one-side test



An example of application

- Sample of $n = 1908$ women, aged 19 to 59 in 1980, who were followed from 1979 to 1985 (source PSID)
- *Response variable and covariates*:
  - ▷ $y_{it} = 1$ if woman i has a job position in year t
  - ▷ age in 1980 (*time-constant*)
  - ▷ race (dummy equal to 1 for a black; *time-constant*)
  - ▷ educational level (number of years of schooling; *time-constant*)
  - ▷ number of children aged 0 to 2 (*time-varying*), aged 3 to 5 (*time-varying*) and aged 6 to 17 (*time-varying*)
  - ▷ permanent income (average income of the husband from 1980 to 1985; *time-constant*)
  - ▷ temporary income (difference between income of the husband in a year and permanent income; *time-varying*)

- Comparison with the results obtained with the JML and MML (based on normal distribution for α_i)
- The main difference is in the estimate of the state dependence effect (γ) that is unreliable under JML; this effect is (likely) overestimated under MML
- As in any fixed-effects approach, the regression coefficients for the time-constant covariates are not estimable; however, the parameter of greatest interest is γ

					pseudo CML			
Parameter	JML	s.e.	MML	s.e.	Basic	s.e.	Improved	s.e.
Kids 0-2	-1.2688	(0.1015)	-0.8832	(0.0825)	-0.7683	(0.1015)	-0.9196	(0.1019)
Kids 3-5	-0.8227	(0.0937)	-0.4390	(0.0736)	-0.4434	(0.0937)	-0.4407	(0.0948)
Kids 6-17	-0.1730	(0.0706)	-0.0819	(0.0393)	-0.0979	(0.0706)	-0.0190	(0.0713)
Temp. inc.	-0.0112	(0.0033)	-0.0036	(0.0030)	-0.0062	(0.0033)	-0.0060	(0.0033)
Lag-res. (γ)	*-0.5696*	(0.0879)	*2.7974*	(0.0653)	*1.6390*	(0.0879)	*1.5660*	(0.0861)

Multivariate extension of the dynamic logit model

- Basic notation:
  - ▷ r: number of response variables observed at each occasion
  - ▷ y_{hit}: categorical response variable h for subject i at occasion t
  - ▷ l_h: number of categories of y_{hit}, indexed from 0 to $l_h - 1$
  - ▷ $\boldsymbol{y}_{it} = \{y_{hit},\ h = 1, \ldots, r\}$: vector of response variables for subject i at occasion t
  - ▷ $\boldsymbol{\alpha}_{it}$: vector of time-varying subject-specific effects
- The conditional distribution of $\boldsymbol{y}_{it}$ given $\boldsymbol{\alpha}_{it}$, $\boldsymbol{x}_{it}$ and $\boldsymbol{y}_{i,t-1}$ is parametrized by marginal (with respect to other response variables) logits and log-odds ratios which may be of *local*, *global* or *continuation* type

- Generalized logits may be of type (with $z = 1, \ldots, l_h - 1$):
  - *local*: $$\log \frac{p(y_{hit} = z | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})}{p(y_{hit} = z-1 | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})}$$
  - *global*: $$\log \frac{p(y_{hit} \geq z | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})}{p(y_{hit} < z | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})}$$
  - *continuation*: $$\log \frac{p(y_{hit} \geq z | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})}{p(y_{hit} = z-1 | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})}.$$
- Global and continuation logits are suitable for ordinal variables; local logits are commonly used with non-ordered categories
- Marginal logits and log-odds ratios are collected in the column vector $\boldsymbol{\eta}(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})$ which has dimension:

$$\sum_h (l_h - 1) + \sum_{h_1 < r} \sum_{h_2 > h_1} (l_{h_1} - 1)(l_{h_2} - 1)$$

- The vector of marginal effects may be expressed as (Gloneck & McCullagh, 1995; Colombi & Forcina, 2001; Bartolucci *et al.*, 2007):

$$\boldsymbol{\eta}(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1}) = \boldsymbol{C} \log[\boldsymbol{M}\boldsymbol{p}(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})]$$

  - ▷ $\boldsymbol{p}(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})$: probability vector for the conditional distribution of $\boldsymbol{y}_{it}$ given $\boldsymbol{\alpha}_{it}$, $\boldsymbol{x}_{it}$ and $\boldsymbol{y}_{i,t-1}$
  - ▷ $\boldsymbol{C}$: matrix of contrasts
  - ▷ $\boldsymbol{M}$: marginalization matrix

- We assume that all the three and higher order log-linear interactions for $\boldsymbol{p}(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})$ are equal to 0, so that the link function is a one-to-one transformation of this probability vector

- A simple Newton algorithm may be used to obtain $\boldsymbol{p}(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})$ from $\boldsymbol{\eta}(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})$

Parametrization of marginal effects

- We then assume that

$$
\begin{aligned}
\boldsymbol{\eta}_1(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1}) &= \boldsymbol{\alpha}_{it} + \boldsymbol{X}_{it}\boldsymbol{\beta} + \boldsymbol{Y}_{it}\boldsymbol{\gamma} \\
\boldsymbol{\eta}_2(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1}) &= \boldsymbol{\phi}
\end{aligned}
$$

  - ▷ $\boldsymbol{\eta}_1(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})$: subvector containing marginal logits
  - ▷ $\boldsymbol{\eta}_2(\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})$: subvector containing marginal log-odds ratios
  - ▷ $\boldsymbol{X}_{it}$: design matrix defined on the basis of $\boldsymbol{x}_{it}$ (e.g. $\boldsymbol{I} \otimes \boldsymbol{x}'_{it}$)
  - ▷ $\boldsymbol{Y}_{it}$: design matrix defined on the basis of $\boldsymbol{y}_{i,t-1}$ (e.g. $\boldsymbol{I} \otimes \boldsymbol{y}'_{i,t-1}$)
  - ▷ $\boldsymbol{\beta}$: vector of regression parameters for the covariates
  - ▷ $\boldsymbol{\gamma}$: vector of parameters for the lagged responses
  - ▷ $\boldsymbol{\phi}$: vector of association parameters

Latent Markov chain

- For each i, the random parameter vectors $\{\boldsymbol{\alpha}_{i1}, \dots, \boldsymbol{\alpha}_{iT}\}$ are assumed to follow a (unobservable) *first-order Markov chain* with
  - ▷ states $\boldsymbol{\xi}_c$, $c = 1, \dots, k$
  - ▷ initial probabilities $\lambda_c(\boldsymbol{y}_{i0})$, $c = 1, \dots, k$
  - ▷ transition probabilities π_{cd}, $c, d = 1, \dots, k$
- Dependence of the initial probabilities on the initial observations in $\boldsymbol{y}_{i0}$ is modelled on the basis of the parametrization

$$\boldsymbol{\psi}(\boldsymbol{y}_{i0}) = \boldsymbol{Y}_{i0}\boldsymbol{\delta}$$

  - ▷ $\boldsymbol{\psi}(\boldsymbol{y}_{i0})$: column vector of logits $\log[\lambda_c(\boldsymbol{y}_{i0})/\lambda_1(\boldsymbol{y}_{i0})]$, $c = 2, \dots, k$
  - ▷ $\boldsymbol{Y}_{i0}$: design matrix depending on $\boldsymbol{y}_{i0}$, typically $\boldsymbol{I} \otimes (1 \quad \boldsymbol{y}'_{i0})$
  - ▷ $\boldsymbol{\delta}$: vector of parameters

Maximum likelihood estimation

- Estimation is performed by maximizing the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_i \log \left[\sum_{\boldsymbol{\alpha}_i} p(\boldsymbol{\alpha}_i) \prod_t p(\boldsymbol{y}_{it} | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1}) \right]$$

- $\ell(\boldsymbol{\theta})$ is maximized by an EM algorithm (Dempster *et al.*, 1977) based on the *complete data log-likelihood*:

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) \;\; = \;\; & \sum_i \Big\{ \sum_c w_{i1c} \log[\lambda_c(\boldsymbol{y}_{i0})] + \sum_c \sum_d z_{icd} \log(\pi_{cd}) + \\ & + \sum_t \sum_c w_{itc} \log[p(\boldsymbol{y}_{it} | \boldsymbol{\xi}_c, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})] \Big\}\end{aligned}$$

  - ▷ w_{itc}: dummy variable equal to 1 if subject i is in latent state c at occasion t (i.e. $\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_c$) and to 0 otherwise
  - ▷ $z_{icd} = \sum_{t>1} w_{i,t-1,c} w_{itd}$: number of transitions from state c to d

EM algorithm

- The EM algorithm performs two steps until convergence in $\ell(\boldsymbol{\theta})$:

  E: compute the conditional expected value of the *complete data log-likelihood* given the current $\boldsymbol{\theta}$ and the observed data

  M: maximize this expected value with respect to $\boldsymbol{\theta}$
- Computing the conditional expected value of $\ell^*(\boldsymbol{\theta})$ is equivalent to computing the conditional expected value of w_{itc} and $w_{i,t-1,c}w_{itd}$. This is done by certain recursions taken from the literature on hidden Markov models (MacDonald & Zucchini, 1997)
- The parameters in $\boldsymbol{\theta}$ are updated at the M-step by simple iterative algorithms. An explicit formula is also available for some parameters

An application (PSID data)

- Dataset concerning $n = 1446$ women followed from 1987 to 1993
- Two binary response variables:
  - ▷ *fertility*: equal to 1 if the woman had given birth to a child
  - ▷ *employment*: equal to 1 if the woman was employed
- Eight covariates (beyond a dummy variable for each year):
  - ▷ *race*: dummy variable equal to 1 for a black woman
  - ▷ *age* in 1986
  - ▷ *education* (in year of schooling)
  - ▷ *child 1-2*: number of children in the family between 1 and 2 years
  - ▷ *child 3-5*, *child 6-13*, *child 14-*
  - ▷ *income of the husband* (in dollars)

- The conditional distribution of $\boldsymbol{y}_{it}$ (containing the two response variables) is modelled by two marginal logits and one log-odds ratio:

$$\log \frac{p(y_{1it}=1|\boldsymbol{\alpha}_{it},\boldsymbol{x}_{it},\boldsymbol{y}_{i,t-1})}{p(y_{1it}=0|\boldsymbol{\alpha}_{it},\boldsymbol{x}_{it},\boldsymbol{y}_{i,t-1})} = \alpha_{1it} + \boldsymbol{x}'_{it}\boldsymbol{\beta}_1 + \boldsymbol{y}'_{i,t-1}\boldsymbol{\gamma}_1$$

$$\log \frac{p(y_{2it}=1|\boldsymbol{\alpha}_{it},\boldsymbol{x}_{it},\boldsymbol{y}_{i,t-1})}{p(y_{2it}=0|\boldsymbol{\alpha}_{it},\boldsymbol{x}_{it},\boldsymbol{y}_{i,t-1})} = \alpha_{2it} + \boldsymbol{x}'_{it}\boldsymbol{\beta}_2 + \boldsymbol{y}'_{i,t-1}\boldsymbol{\gamma}_2$$

$$\log \frac{p(y_{1it}=1,y_{2it}=1|\boldsymbol{\alpha}_{it},\boldsymbol{x}_{it},\boldsymbol{y}_{i,t-1})}{p(y_{1it}=1,y_{2it}=0|\boldsymbol{\alpha}_{it},\boldsymbol{x}_{it},\boldsymbol{y}_{i,t-1})} + \\ + \log \frac{p(y_{1it}=0,y_{2it}=0|\boldsymbol{\alpha}_{it},\boldsymbol{x}_{it},\boldsymbol{y}_{i,t-1})}{p(y_{1it}=0,y_{2it}=1|\boldsymbol{\alpha}_{it},\boldsymbol{x}_{it},\boldsymbol{y}_{i,t-1})} = \delta$$

- The process $\{\boldsymbol{\alpha}_{i1},\ldots,\boldsymbol{\alpha}_{iT}\}$, with $\boldsymbol{\alpha}_{it}=(\alpha_{1it},\alpha_{2it})'$, follows an homogenous Markov chain with initial probabilities depending on $\boldsymbol{y}_{i0}$

- The model is estimated for an increasing number of latent states (k) from 1 to 5, with the fit measured on the basis of:

$$
\begin{aligned}
AIC &= -2\ell(\hat{\boldsymbol{\theta}}) + 2g \\
BIC &= -2\ell(\hat{\boldsymbol{\theta}}) + g\log(n)
\end{aligned}
$$

  - g: number of parameters

	k				
	1	2	**3**	4	5
log-lik.	-6219.0	-6050.0	**-6011.5**	-6004.7	-5993.6
# par.	37	44	**53**	64	77
AIC	12512	12188	**12129**	12137	12141
BIC	12707	12420	**12409**	12475	12548

- The results lead us to select $k = 3$ latent states

- Estimates of the regression parameters ($k = 3$):

Effect	**logit fertility**	**logit employment**	**log-odds ratio**
intercept (average of support points)	-2.310	2.285	-1.238
race	-0.238	0.175	-
age	-0.219	0.060	-
age$^2/100$	-1.117	-0.110	-
education	0.158	0.085	-
child 1-2	0.180	-0.113	-
child 3-5	-0.382	-0.168	-
child 6-13	-0.613	0.027	-
child 14-	-0.891	0.071	-
income of the husband/1000	0.002	-0.011	-
lagged fertility	-1.476	-0.726	-
lagged employment	0.321	0.970	-

- Negative state dependence for fertility, positive state dependence for employment and negative association between the response variables

- Estimated initial probability vector and transition probability matrix (averaged over all the subjects in the sample)

$$\hat{\boldsymbol{\lambda}} = \begin{pmatrix} 0.010 \\ 0.266 \\ 0.634 \end{pmatrix}, \qquad \hat{\boldsymbol{\Pi}} = \begin{pmatrix} 0.947 & 0.050 & 0.003 \\ 0.068 & 0.888 & 0.044 \\ 0.003 & 0.092 & 0.906 \end{pmatrix}$$

- The hypothesis that the transition matrix is diagonal must be rejected with a likelihood ratio statistic equal to 32.079
- The assumption that the parameters in $\boldsymbol{\alpha}_{it}$ for the unobserved heterogeneity are time-constant is restrictive; under this hypothesis the estimates of the association parameters are considerably different (e.g. 1.791 vs. 0.970 for the state dependence on employment)

Conclusions

- The proposed pseudo CML estimator is *very simple to use* and does not require to formulate *any assumption* on the distribution of the subject-specific effects
- The estimator is only consistent when $\gamma_0 = 0$, but simulation results show that its *bias* is very limited even when $\gamma_0 \neq 0$
- With respect to the *HK estimator* (a benchmark in this field):
  - ▷ it shows a clear advantage in terms of efficiency due to the larger actual sample size
  - ▷ can be used with $T \geq 2$ (instead of $T \geq 3$) and witout restrictions on the covariates structures (even with time dummies)
- By exploiting the proposed sandwich estimator for computing $s.e.(\hat{\boldsymbol{\theta}})$, we can simply test the *hypothesis of absence of state dependence*

- As in any fixed-effects approach, it is not possible to estimate the effect of *time-varying covariates*, however:
  - ▷ the parameter of greatest interest is usually γ (state dependence)
  - ▷ the approach can be combined with an MML approach
- It seems possible to exploit the approach for *other fixed-effects models* in which there are no sufficient statistics for the subject-specific parameters α_i
- Examples are *extensions of the Rasch (1961) model* in which:
  - ▷ the responses are allowed to be dependent even conditionally on α_i
  - ▷ a more complex parametrization of the probability of success is used (2PL-model with discriminant index)

- The *multivariate extension* of the dynamic logit model leads to a flexible class of models which may be used with ordinal and non-ordinal categorical variables and extend the LM model
- The approach allows to model the *contemporary association* between the response variables
- Modeling the vector of subject-specific parameters by a *latent Markov chain* allows us to take into account that the effect of unobservable covariates may be not time-constant
- The *EM algorithm* may be efficiently implemented by recursions taken from the hidden Markov literature
- Special attention has to be payed to the *multimodality of the model likelihood*, e.g. by adopting random starting strategies

Main references

Akaike, H. (1973), Information theory and an extension of the likelihood principle, in *Proceedings of the Second International Symposium of Information Theory*, Ed. Petrov, B.N. & Csaki, F. Budapest: Akademiai Kiado.

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

Bartolucci, F. & Farcomeni, A. (2008), A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure, *Journal of the American Statistical Association, in press*.

Bartolucci, F. & Nigro, V. (2008), Approximate conditional inference for panel logit models allowing for state dependence and unobserved heterogeneity, *Technical report*, arXiv:math/0702774.

Bartolucci, F. & Nigro, V. (2007), A Dynamic Model for Binary Panel Data With Unobserved Heterogeneity Admitting a Root-N Consistent Conditional Estimator, *CEIS Working Paper No. 98*.

Bartolucci, F., Colombi, R. & Forcina, A. (2007), An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints, *Statistica Sinica*, **17**, 691–711.

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. & Singer B. Cambridge: Cambridge University Press.

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

Colombi R. & Forcina A. (2001), Marginal regression models for the analysis of positive association of ordinal response variables, *Biometrika*, **88**, 1007–1019.

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

Cox, D. R. & Wermuth, N. (1994), A note on the quadratic exponential binary distribution, *Biometrika*, **81**, 403–408.

Dempster A. P., Laird, N. M. & Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

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

Glonek, G. F. V. & McCullagh, P., (1995), Multivariate Logistic Models, *Journal of the Royal Statistical Society B*, **57**, 533-546.

Gourieroux, C. & Monfort, A. (1993), Pseudo-likelihoods methods, in *Handbook of Statistics*, Vol. 11, Ed. Maddala G.S., Rao C. R. & Vinod H.D. Elsevier Science, North Holland.

Hahn, J. & Kuersteiner, G. (2004), Bias reduction for dynamic nonlinear panel models with fixed effects, unpublished manuscript.

Hahn, J. & Newey, W.K. (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. & 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. & Manski C. A. Cambridge, MA: MIT Press.

Heckman, J.J. (1981c), The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, in *Structural Analysis of discrete data*, Cambridge, MA.: MIT press.

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

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.

Lazarsfeld, P. F. & Henry, N. W. (1968), *Latent Structure Analysis*, Houghton Mifflin, Boston.

MacDonald, I. L. & Zucchini, W. (1997), *Hidden Markov and other Models for Discrete-Valued Time Series*, Chapman and Hall, London.

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

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

Newey W. K. & McFadden D. L. (1994), Large Sample Estimation and Hypothesis Testing", in *Handbook of Econometrics*, Vol. 4, Ed. Engle R. F. & McFadden D. L. Amsterdam: North-Holland.

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

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.

Ten Have, T. R. & Morabia, A. (1999), Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data, *Biometrics*, **55**, pp. 85-93.

Todem, D., Kim, K. & Lesaffre, E. (2007), Latent-variable models for longitudinal data with bivariate ordinal outcomes, *Statistics in Medicine*, **26**, 1034–1054.

Vermunt, J. K., Langeheine, R. & Böckenholt, U. (1999), Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates, *Journal of Educational and Behavioral Statistics*, **24**, 179–207.

White, H. (1982): "Maximum likelihood estimation of misspecified models", *Econometrica*, 50, 1–26.

Wiggins, L. M. (1973), *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*, Amsterdam: Elsevier.

Wooldridge, J. M. (2000), A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables, *Economics Letters*, **68**, 245-250.