mixmcm: a Stata command for estimating mixture of Markov chain models using ML and the EM algorithm

Legrand D. F. SAINT-CYR and Laurent PIET

SMART, AGROCAMPUS OUEST, INRA, 35000, Rennes, France

French Stata Users Group meeting
Paris | July 6

Background

- Markov chain model (MCM) is a widely used modelling approach in several strands of the literature
  - MCM enables analysing dynamic stochastic process within a given population (future states depend on the past accordingly to some probabilities)
  - Numerous applications in economics (firm dynamics, unemployment, ...), medicine (illness treatment), sociology (population mobility), ...

Background

- Markov chain model (MCM) is a widely used modelling approach in several strands of the literature
  - MCM enables analysing dynamic stochastic process within a given population (future states depend on the past accordingly to some probabilities)
  - Numerous applications in economics (firm dynamics, unemployment, ...), medicine (illness treatment), sociology (population mobility), ...
- Heterogeneous behaviours in several cases that are generally unobserved and cannot be captured by observable agent characteristics

Background

- Markov chain model (MCM) is a widely used modelling approach in several strands of the literature
  - MCM enables analysing dynamic stochastic process within a given population (future states depend on the past accordingly to some probabilities)
  - Numerous applications in economics (firm dynamics, unemployment, ...), medicine (illness treatment), sociology (population mobility), ...
- Heterogeneous behaviours in several cases that are generally unobserved and cannot be captured by observable agent characteristics
- Some available commands in STATA allow estimating finite mixture models to capture unobserved heterogeneity
  - Official commands (*fmm*)
  - Users written commands (*gllamm*, *lclogit*, ...)
- **Impossible to estimate directly a mixture of Markov chain models using the available commands in STATA**

The mixed Markov chain model (MMCM)

- MMCM describes the dynamics of N agents on a finite state space K over a time period T with heterogeneous transition processes

- Probability density function of MCM

$$f(\mathbf{y}_i) = \prod_{t=1}^{T_i} P(y_{it} = k | y_{it-1} = j), \qquad \forall i \in N; \quad \forall j, k \in K$$

$\mathbf{y}_i = (y_{i0}, y_{i1}, \cdots, y_{iT_i}); \quad T_i \leq T$

The mixed Markov chain model (MMCM)

- MMCM describes the dynamics of N agents on a finite state space K over a time period T with heterogeneous transition processes

- Probability density function of MCM

$$f(\mathbf{y}_i) = \prod_{t=1}^{T_i} P(y_{it} = k | y_{it-1} = j), \qquad \forall i \in N; \quad \forall j, k \in K$$

$\mathbf{y}_i = (y_{i0}, y_{i1}, \cdots, y_{iT_i}); \quad T_i \leq T$

- Probability density function of MMCM

$$f(\mathbf{y}_i) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{y}_i)$$

$0 \leq \pi_g \leq 1$: probability of belonging to type g

The model specification

- Multinomial logit specification of transition probabilities

$$P(y_{it} = k | y_{it-1} = j, g) = \frac{\exp(\boldsymbol{\beta}'_{jk|g} \mathbf{x}_{it-1})}{\sum_{l=1}^{K} \exp(\boldsymbol{\beta}'_{jl|g} \mathbf{x}_{it-1})}$$

$\boldsymbol{\beta}_{jj|g} = 0$ for identification

The model specification

- Multinomial logit specification of transition probabilities

$$P(y_{it} = k | y_{it-1} = j, g) = \frac{\exp(\boldsymbol{\beta}'_{jk|g} \mathbf{x}_{it-1})}{\sum_{l=1}^{K} \exp(\boldsymbol{\beta}'_{jl|g} \mathbf{x}_{it-1})}$$

$\boldsymbol{\beta}_{jj|g} = 0$ **for identification**

- Probability of type membership
  - Fractional multinomial logit for parametric specification

$$P(g_i = g | \mathbf{z}_i) = \frac{\exp(\boldsymbol{\lambda}'_g \mathbf{z}_i)}{\sum_{h=1}^{G} \exp(\boldsymbol{\lambda}'_h \mathbf{z}_i)} \quad (\forall g \in G - 1)$$

  - Non-parametric estimation implies that $P(g_i = g)$ are the same for all agents

The EM algorithm under incomplete information

- E-step: Compute the probability of belonging to type g

$$v_{i|g}^{(p+1)} = \frac{P(g_i = g)^{(p)} \prod_{t=1}^{T_i} \prod_{j,k}^{K} \left[P(\mathbf{x}_{it-1}; \boldsymbol{\beta}_{jk|g}^{(p)}) \right]^{d_{ijkt}}}{\sum_{h=1}^{G} P(h_i = h)^{(p)} \prod_{t=1}^{T_i} \prod_{j,k}^{K} \left[P(\mathbf{x}_{it-1}; \boldsymbol{\beta}_{jk|h}^{(p)}) \right]^{d_{ijkt}}}$$

$d_{ijkt} = 1$ if agent i move from j to k

The EM algorithm under incomplete information

- E-step: Compute the probability of belonging to type g

$$v_{i|g}^{(p+1)} = \frac{P(g_i = g)^{(p)} \prod_{t=1}^{T_i} \prod_{j,k}^{K} \left[P(\mathbf{x}_{it-1}; \boldsymbol{\beta}_{jk|g}^{(p)})\right]^{d_{ijkt}}}{\sum_{h=1}^{G} P(h_i = h)^{(p)} \prod_{t=1}^{T_i} \prod_{j,k}^{K} \left[P(\mathbf{x}_{it-1}; \boldsymbol{\beta}_{jk|h}^{(p)})\right]^{d_{ijkt}}}$$

$d_{ijkt} = 1$ if agent i move from j to k

- M-step: Maximize the conditional log-likelihood
  - Parameters of the transition probabilities

$$\boldsymbol{\beta}^{(p+1)} = argmax_{\boldsymbol{\beta}} \sum_{i=1}^{N} \sum_{g=1}^{G} v_{i|g}^{(p+1)} \sum_{t=1}^{T_i} \sum_{j,k}^{K} d_{ijkt} \ln\left[P(\mathbf{x}_{it-1}; \boldsymbol{\beta}_{jk|g})\right]$$

  - Parameters of the mixing distribution
    - Parametrically: $\boldsymbol{\lambda}^{(p+1)} = argmax_{\boldsymbol{\lambda}} \sum_{i=1}^{N} \sum_{g=1}^{G} v_{i|g}^{(p+1)} \ln[P(\mathbf{z}_i; \boldsymbol{\lambda}_g)]$
    - Non-parametrically: $\pi_g^{(p+1)} = \frac{\sum_{i=1}^{N} v_{i|g}^{(p+1)}}{\sum_{i=1}^{N} \sum_{h=1}^{G} v_{i|h}^{(p+1)}}$

The mixmcm command

- The generic syntax for mixmcm:

  mixmcm *depvar* [*indepvars*] [*if*] [*in*] [*weight*], **id**(*varname*) **time**var(*varname*) [*options*]

- The options for mixmcm:
  - * **id**(*varname*): numeric variable identifying agents
  - * **timevar**(*varname*): numeric variable identifying time

The mixmcm command

- The generic syntax for mixmcm:

  mixmcm *depvar* [*indepvars*] [*if*] [*in*] [*weight*], **id**(*varname*) **timevar**(*varname*) [*options*]

- The options for mixmcm:
  - * **id**(*varname*): numeric variable identifying agents
  - * **timevar**(*varname*): numeric variable identifying time
  - **ncomponents**(*#1 #2, selcrit(name) graph(namelist, twoway_ options) force save(filename, replace detail)*)
  - **membership**(*varlist, fmlogit_ options*)
  - **emiterate**(*lr(#1 #2, eps) sr(#1 #2) seed(numlist) emlog)*)
  - **noconstant**: suppress constant term in the specification of transition probabilities
  - **constraints**(*p_ #component_ initialstate_ finalstate*)

Application

- Data come from the free online version of RICA the French implementation of the FADN on commercial farms from 2000 to 2010
- Some modifications of the data to identify Markov states and to generate some explanatory variables

Application

- Data come from the free online version of RICA the French implementation of the FADN on commercial farms from 2000 to 2010
- Some modifications of the data to identify Markov states and to generate some explanatory variables
- The ten first lines of the dataset

idnum	year	ebexp(€)	subex(€)	debt(%)	education	corporate	category
963	2000	36804.39	19798.40	5.35	1	0	medium
963	2001	28861.00	23290.00	5.35	1	0	medium
963	2002	30000.12	25990.33	5.35	1	0	medium
963	2003	5159.31	17527.58	5.35	1	0	medium
1525	2006	58895.00	17542.00	20.40	1	1	large
1525	2007	51726.00	16284.00	20.40	1	1	verylarge
1525	2008	54940.00	26491.00	20.40	1	1	verylarge
1525	2009	51883.00	16015.00	20.40	1	1	verylarge
1525	2010	88685.00	14900.00	20.40	1	1	verylarge
1534	2006	90051.00	78402.00	47.90	1	1	verylarge

Output results

- STATA procedure for estimating the MMCM using RICA French farm dataset

```
. use mixmcmdata.dta, clear

. constraint 1 p_*_medium_verylarge = 0
. constraint 2 p_*_verylarge_medium = 0

. mixmcm category corporate ebexp subex, id(idnum) time(year)
  nc(1 4, selcrit(aic3) graph(aic bic caic aic3, ytitle("Criteria") force save(ictable, replace detail))
  members(education debt, baseoutcome(_proba_1)) em(lr(10 100, 0.0001) sr(5 5)) const(1 2)
```

Output results

- STATA procedure for estimating the MMCM using RICA French farm dataset

```
. use mixmcmdata.dta, clear

. constraint 1 p_*_medium_verylarge = 0
. constraint 2 p_*_verylarge_medium = 0

. mixmcm category corporate ebexp subex, id(idnum) time(year)
  nc(1 4, selcrit(aic3) graph(aic bic caic aic3, ytitle("Criteria") force save(ictable, replace detail))
  members(education debt, baseoutcome(_proba_1)) em(lr(10 100, 0.0001) sr(5 5)) const(1 2)
```

- Selection of the optimal number of components using information criteria

[Graphic]

Output results

- STATA procedure for estimating the MMCM using RICA French farm dataset

```
. use mixmcmdata.dta, clear

. constraint 1 p_*_medium_verylarge = 0
. constraint 2 p_*_verylarge_medium = 0

. mixmcm category corporate ebexp subex, id(idnum) time(year)
  nc(1 4, selcrit(aic3) graph(aic bic caic aic3, ytitle("Criteria") force save(ictable, replace detail))
  members(education debt, baseoutcome(_proba_1)) em(lr(10 100, 0.0001) sr(5 5)) const(1 2)
```

- Selection of the optimal number of components using information criteria

[Graphic]

- Transition probabilities and type membership parameters

[Table1]

Output results

- STATA procedure for estimating the MMCM using RICA French farm dataset

```
. use mixmcmdata.dta, clear

. constraint 1 p_*_medium_verylarge = 0
. constraint 2 p_*_verylarge_medium = 0

. mixmcm category corporate ebexp subex, id(idnum) time(year)
  nc(1 4, selcrit(aic3) graph(aic bic caic aic3, ytitle("Criteria") force save(ictable, replace detail))
  members(education debt, baseoutcome(_proba_1)) em(lr(10 100, 0.0001) sr(5 5)) const(1 2)
```

- Selection of the optimal number of components using information criteria

  [Graphic]

- Transition probabilities and type membership parameters

  [Table1]

- Results stored in e() and saved if specified by the user

  [logfile] [Table2]

Future steps

- Adapt constraints to enable estimating the mover-stayer model (with several mover types)
- Allow for different parametric forms for the mixing distribution (logit, poisson, ...)
- Enable ***mixmcm*** accounting for new entries and exits in the population under study and estimating their parameters
- Write postestimation commands for ***mixmcm***:
  - predict transition probabilities, margins for transition and membership explanatory variables
  - perform projections of the population distribution across the states of space and over time

Thank you!

legrand.saint-cyr@inra.fr