Latent drop-out classes in linear quantile hidden Markov models

M.F. Marino[1] N. Tzavidis[2] M. Alfó[3]

[1]University of Perugia
[2]University of Southampton
[3]Sapienza, University of Rome

Unversity of Pisa

Recent Advances in Quantile and M-quantile Regression

July 15, 2015

Longitudinal data

- Data are repeatedly collected over time on a sample of units
- We have a two stage sample

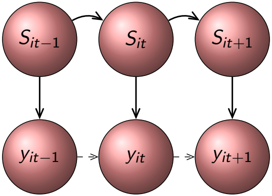
$$[\mathbf{y}_i, \mathbf{X}_i] = \Big([y_{i1}, \mathbf{x}_{i1}], ..., [y_{it}, \mathbf{x}_{it}], ..., [y_{iT}, \mathbf{x}_i]\Big)$$

  - y_{it}'s are realizations of *continuous* random variables Y_{it}
  - $\mathbf{x}_{it}$'s are vectors of p explanatory variables

Observations coming from the same individual are associated because of the presence of unobserved factors (unobserved heterogeneity)

Hidden Markov models for longitudinal data *(Bartolucci et al. 2012)*

Unobserved dynamics are captured via random parameters evolving over time according to a homogeneous, first order, hidden Markov chain $\{S_{it}\}$



- For a given $t = 1, ..., T$, the outcome y_{it} is influenced by S_{it} only
- Conditional on the hidden states, longitudinal observations are independent

$$f_{y|s}(\mathbf{y}_i \mid \mathbf{s}_i) = \prod_{t=1}^{T} f_{y|s}\left(y_{it} \mid s_{it}\right)$$

Linear quantile hidden Markov models - lqHMM *(Farcomeni, 2012)*

AIM: Analyse the relation between a set of explanatory variables and the quantiles of a **continuous** outcome

Conditional on a quantile-specific hidden Markov chain, the τ-th (conditional) quantile regression model is defined by

$$Q_\tau(y_{it} \mid s_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta}(\tau) + \mathbf{w}'_{it}\boldsymbol{\alpha}_{s_{it}(\tau)}$$

ML estimates can be obtained by *conveniently* assuming a (conditional) asymmetric Laplace distribution *(Geraci and Bottai, 2007)*

$$\text{ALD}\left(\mathbf{x}'_{it}\boldsymbol{\beta}(\tau) + \mathbf{w}'_{it}\boldsymbol{\alpha}_{s_{it}(\tau)}, \sigma, \tau\right)$$

- Let the longitudinal study be designed to collect T repeated measures of a **continuous** response variable

$$\mathbf{y}_i = (y_{i1}, ..., y_{iT})$$

- Some units drop-out before the end of the study

$$\mathbf{y}_i = (\mathbf{y}_i^o, \mathbf{y}_i^m) = (y_{i1}, ..., y_{iT_i}, NA, ..., NA)$$

Missing data generating process

- IGNORABLE: Conditional on $(\mathbf{y}_i^o, \mathbf{x}_i)$ the missing data process does not provide information on the missing responses
- **NON-IGNORABLE:** the probability that a unit remains into the study depends on unobserved responses

"Joint" models for the observed and the missing data process are often considered in this context *(Little and Rubin, 2002)*

Modeling non-ignorable drop-out

- Selection models (Heckman, 1976)

$$f_{y,t}(\mathbf{y}_i, T_i) = f_y(\mathbf{y}_i) f_{t|y}(T_i \mid \mathbf{y}_i)$$

- Pattern mixture models (Little, 1993)

$$f_{y,t}(\mathbf{y}_i, T_i) = f_t(T_i) f_{y|t}(\mathbf{y}_i \mid T_i)$$

- Random coefficient based missing data models

$$f_{y,t}(\mathbf{y}_i, T_i) = \int f_{t|u}(T_i \mid \mathbf{u}_i) f_{y|b}(\mathbf{y}_i \mid \mathbf{b}_i) dF_{u,b}(\mathbf{u}_i, \mathbf{b}_i)$$

  - If $\mathbf{u}_i = \mathbf{b}_i \rightarrow$ shared parameter models (Wu and Carroll, 1988)
  - If a survival model describes the time to drop-out $\rightarrow$ joint models (Rizopoulos, 2012)

Pattern mixture models (PMMs - Little, 1993)

- Each individual has its own propensity to drop-out from the study
- Individuals with similar drop-out history share similar (unobserved) features
- The model for the whole population is given by a mixture over drop-out patterns
- PMMs are weakly identifiable due a (potentially) large number of patterns $\rightarrow$ identifiability constraints are needed

Latent drop-out class model *(Roy, 2003; Roy and Daniels, 2008)*

- Individual propensities to drop-out from the study can be described by a latent drop-out (LDO) class variable with G ordered categories
- The length of the observation window influences the probability of belonging to one of the G LDO classes
- Conditional on the LDO class variable, the observed and the missing data process are independent

LDO classes in linear quantile HMMs *(Marino et al., 2015)*

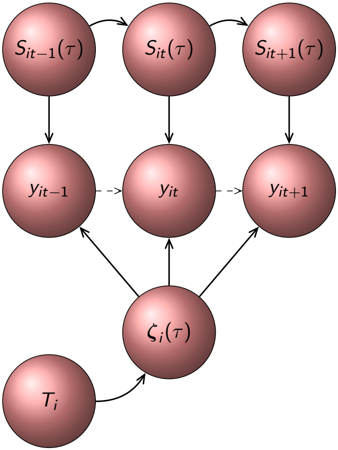
We extend the proposal by Farcomeni (2012) in a LDO class perspective

- *Quantile regression offers a* **complete picture** *of the outcome distribution and ensures* **robustness** *against potential outliers.*
- *The hidden Markov structure allows for time-varying dependence*
- *LDO classes help account for potentially* **non-ignorable drop-outs**

For a given quantile $\tau \in (0, 1)$

- let $\{S_{it}(\tau)\}$ be a quantile-specific, homogeneous, hidden Markov chain
- let $\boldsymbol{\zeta}_i(\tau) = (\zeta_{i1}(\tau), ..., \zeta_{iG}(\tau))$ be a quantile-dependent LDO class membership

Linear quantile HMM+LDO: model assumptions



- Latent variables $\zeta_i(\tau)$ and $S_{it}(\tau)$ are independent
- For a given time occasion, y_{it} is influenced only by $S_{it}(\tau)$ and $\zeta_i(\tau)$
- Conditional on the latent variables, longitudinal observations are independent

$$f_y(\mathbf{y}_i \mid \mathbf{s}_i, \zeta_i; \tau) = \prod_{t=1}^{T_i} f_y\left(y_{it} \mid s_{it}, \zeta_i; \tau\right)$$

- Conditional on ζ_i, the observed and the missing data process are independent

- Model for the LDO class variable

$$\Pr\left(\sum_{l=1}^{g} \zeta_{il} = 1 \mid T_i\right) = \frac{\exp\{\lambda_{0g} + \lambda_1 T_i\}}{1 + \exp\{\lambda_{0g} + \lambda_1 T_i\}}$$

- Model for the hidden Markov chain

$$f(\mathbf{s}_i) = \delta_{s_{i1}} \prod_{t=2}^{T_i} q_{s_{it-1} s_{it}} \quad i = 1, .., n$$

- Conditional (on ζ and S_{it}) model for the **complete** longitudinal responses

$$[Y_{it} \mid S_{it} = s_{it}, \zeta_{ig} = 1; \tau] \sim ALD\left(\mathbf{x}_{it}'\beta(\tau) + \mathbf{z}_{it}'\mathbf{b}_g(\tau) + \mathbf{w}_{it}'\boldsymbol{\alpha}_{s_{it}(\tau)}, \, \sigma, \, \tau\right)$$

Linear quantile HMM+LDO: the likelihood

The individual contribution to the observed (conditional) data likelihood is

$$L_i(\cdot \mid T_i; \tau) = \int \sum_{g=1}^{G} \sum_{\mathbf{s}_i(\tau)} \left\{ \prod_{t=1}^{T} f_{y|sb}(y_{it} \mid s_{it}, \mathbf{b}_g; \tau) \delta_{s_{i1}}(\tau) \prod_{t=2}^{T} q_{s_{it-1} s_{it(\tau)}}(\tau) \right\} \pi_{ig}(T_i; \tau) \, d\mathbf{y}_i^m \tag{1}$$

The LDO class variable summarizes the information on the dependence between $\mathbf{y}_i$ and T_i; missing data can be integrated out from equation (1)

$$L_i(\cdot \mid T_i; \tau) = \prod_{i=1}^{n} \sum_{g=1}^{G} \sum_{\mathbf{s}_i(\tau)} \left\{ \prod_{t=1}^{T_i} f_{y|sb}(y_{it}^o \mid s_{it}, \mathbf{b}_g; \tau) \delta_{s_{i1}}(\tau) \prod_{t=2}^{T_i} q_{s_{it-1} s_{it(\tau)}}(\tau) \right\} \pi_{ig}(T_i; \tau) \tag{2}$$

and inference can be based on the observed data only

Parameter estimation, inference and model selection

- An EM algorithm (Dempster et al., 1977) may be used to derive parameter estimates
- Extended forward and backward variables (Baum et al., 1970) can be exploited to simplify the computation
- Confidence intervals for parameter estimates are obtained via a non-parametric block bootstrap (Lahiri, 1999)
- The number of LDO classes and hidden states are treated as known and estimated via model selection techniques

Application: the CD4 dataset

- AIM: analysing HIV progression over time via the count of CD4 cells
- 369 men affected by HIV are observed for 1 to 12 occasions
- CD4 count levels are measured at each visit
- The following covariates are measured
  - *Age*: age at seroconversion (centred at 30)
  - *Drugs*: drug use
  - *Packs*: packs of cigarette per day
  - *Partners*: number of sexual partners
  - *CESD*: depression symptoms measured according to the CESD scale
  - $Time_{sero}$: years since seroconversion

We model the quantiles of the log-transformed CD4 counts and compare results obtained under lqHMM and lqHMM+LDO

We focus on

- State-dependent intercept
- LDO-dependent slope for $\text{Time}_{\text{sero}}$

Fixed and state-dependent parameters for the median

	lqHMM		lqHMM+LDO	
# Par	36		33	
Log-L	-1082.530		-1018.042	
BIC	2377.802		**2231.139**	
α_1	5.628	(5.074; 5.753)	6.043	(5.931; 6.114)
α_2	6.198	(6.014; 6.252)	6.416	(6.323; 6.502)
α_3	6.524	(6.393; 6.574)	6.719	(6.647; 6.825)
α_4	6.805	(6.719; 6.874)	7.040	(6.973; 7.215)
α_5	7.191	(7.084; 7.291)	-	- -
Age	-0.003	(-0.007;0.005)	0.004	(-0.001; 0.007)
Drugs	0.036	(-0.016; 0.110)	0.072	(-0.006; 0.145)
Packs	0.049	(0.014; 0.068)	0.042	(0.014; 0.054)
Partners	0.002	(-0.003 ; 0.012)	0.005	(0.000; 0.012)
CESD	-0.005	(-0.007;-0.001)	-0.004	-(-0.006 -0.002)
$\text{Time}_{\text{sero}}$	-0.110	(-0.126; -0.084)	-0.146	(-0.175; -0.119)

- Under lqHMM, a further hidden state is needed
- State-specific intercepts identify increasing CD4 count levels
- Packs of cigarettes and number of sexual partners have a positive effect, while age and drug use play no role. More sever depression symptoms lead to decreasing CD4 counts
- CD4 counts decrease as the time since seroconversion increases

LDO class parameters

In the longitudinal data model

b_1	b_2	b_3	b_4
-0.497 (-0.667 -0.452)	-0.176 (-0.200 -0.155)	-0.070 (-0.098 -0.056)	0.033 (-0.023 0.047)

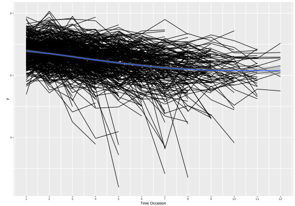
The decrease in CD4 counts over time progressively reduces when moving towards higher LDO classes

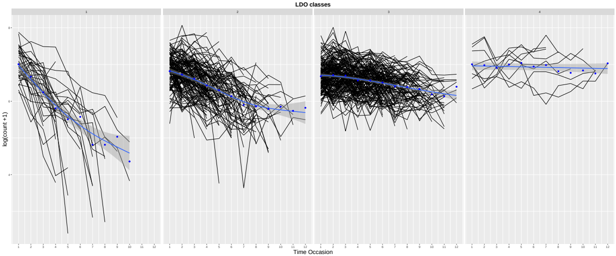
In the LDO class model

λ_{01}	λ_{02}	λ_{03}	λ_1
-1.062 (-2.112; -0.241)	1.113 (0.013; 2.102)	4.089 (2.002; 5.299)	-0.193 (-0.318; -0.065)

When the length of the observation window increases, the probability of "higher" categories increases

Individual trajectories





Concluding remarks

- Sources of unobserved heterogeneity are modelled via a hidden Markov chain
- Bias in the parameter estimates is avoided considering the LDO class variable
- Clustering of units in homogeneous LDO classes offers a clearer interpretation of results
- The semi-parametric nature of the latent variables ensures model flexibility

Basic References

Bartolucci, F., Farcomeni, A., and Pennoni, F. *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. Taylor & Francis, 2012.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, pages 164–171, 1970.

Dempster, A., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

Farcomeni, A. Quantile regression for longitudinal data based on latent Markov subject-specific parameters. *Statistics and Computing*, 22, 2012.

Geraci, M. and Bottai, M. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8 (1):140–54, 2007.

Heckman, J. J. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER, 1976.

Lahiri, S. N. Theoretical comparisons of block bootstrap methods. *Annals of Statistics*, pages 386–404, 1999.

Little, R. J. A. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88 (421):125–134, 1993.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*. Wiley, 2002.

Marino, M. F., Tzavidis, N., and Alfo, M. Quantile regression for longitudinal data: unobserved heterogeneity and informative missingness. *arXiv preprint arXiv:1501.02157*, 2015.

Rizopoulos, D. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis*, 56(3):491–501, 2012.

Roy, J. Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, 59(4):829–836, 2003.

Roy, J. and Daniels, M. J. A general class of pattern mixture models for nonignorable dropout with many possible dropout times. *Biometrics*, 64(2):538–545, 2008.

Wu, M. C. and Carroll, R. J. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188, 1988.