Latent drop-out classes in linear quantile hidden Markov models

M.F. Marino\textsuperscript{1} N. Tzavidis\textsuperscript{2} M. Alfó\textsuperscript{3}

\textsuperscript{1}University of Perugia
\textsuperscript{2}University of Southampton
\textsuperscript{3}Sapienza, University of Rome

University 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

\[ y_{it}, X_{i} \] = \left( [y_{11}, x_{11}], \ldots, [y_{it}, x_{it}], \ldots, [y_{iT}, x_{i}] \right)

- \( y_{it} \)'s are realizations of continuous random variables \( Y_{it} \)
- \( x_{it} \)'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

\[
\begin{align*}
    f_{y|s}(y_i | s_i) &= \prod_{t=1}^{T} f_{y|s}(y_{it} | s_{it})
\end{align*}
\]
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 $\tau$-th (conditional) quantile regression model is defined by

$$Q_{\tau}(y_{it} \mid s_{it}) = x_{it}'\beta(\tau) + w_{it}'\alpha_{s_{it}(\tau)}$$

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

$$\text{ALD} \left( x_{it}'\beta(\tau) + w_{it}'\alpha_{s_{it}(\tau)}, \sigma, \tau \right)$$
Drop-out in longitudinal studies

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

$$y_i = (y_{i1}, \ldots, y_{iT})$$

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

$$y_i = (y_{i}^{o}, y_{i}^{m}) = (y_{i1}, \ldots, y_{iT}, \text{NA}, \ldots, \text{NA})$$

**Missing data generating process**

- **IGNORABLE**: Conditional on $(y_{i}^{o}, x_i)$ 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)
  \[ f_{y,t}(y_i, T_i) = f_y(y_i)f_{t|y}(T_i | y_i) \]

- Pattern mixture models (Little, 1993)
  \[ f_{y,t}(y_i, T_i) = f_t(T_i)f_{y|t}(y_i | T_i) \]

- Random coefficient based missing data models
  \[ f_{y,t}(y_i, T_i) = \int f_{t|u}(T_i | u_i)f_{y|b}(y_i | b_i) dF_{u,b}(u_i, b_i) \]

- If \( u_i = b_i \) → shared parameter models (Wu and Carroll, 1988)
- If a survival model describes the time to drop-out → 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 to a (potentially) large number of patterns → identifiability constraints are needed
Latent drop-out class model \cite{Roy2003,Roy2008}

- 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.
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 $\zeta_i(\tau) = (\zeta_{i1}(\tau), \ldots, \zeta_{iG}(\tau))$ 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

$$f_y(y_i \mid s_i, \zeta_i; \tau) = \prod_{t=1}^{T_i} f_y(y_{it} \mid s_{it}, \zeta_i; \tau)$$

- Conditional on $\zeta_i$, the observed and the missing data process are independent
Model for the LDO class variable

\[ \Pr \left( \sum_{l=1}^{g} \zeta_{il} = 1 \mid T_i \right) = \frac{\exp\{\lambda_0 g + \lambda_1 T_i\}}{1 + \exp\{\lambda_0 g + \lambda_1 T_i\}} \]

Model for the hidden Markov chain

\[ f(s_i) = \delta_{s_{i1}} \prod_{t=2}^{T_i} q_{s_{it-1} s_{it}} \quad i = 1, \ldots, n \]

Conditional (on \( \zeta \) and \( S_{it} \)) model for the complete longitudinal responses

\[ [Y_{it} \mid S_{it} = s_{it}, \zeta_{ig} = 1; \tau] \sim ALD \left( x_{it}' \beta(\tau) + z_{it}' b_g(\tau) + w_{it}' \alpha_{s_{it}(\tau)}, \sigma, \tau \right) \]
The individual contribution to the observed (conditional) data likelihood is

\[ L_i(\cdot \mid T_i; \tau) = \int \sum_{g=1}^{G} \sum_{s_i(\tau)} \left\{ \prod_{t=1}^{T_i} f_{y|s_s}(y_{it} \mid s_{it}, b_g; \tau) \delta_{s_1}^{(\tau)} \prod_{t=2}^{T_i} q_{s_{it-1}s_{it}(\tau)}(\tau) \right\} \pi_{ig}(T_i; \tau) \, dy_i^m \]

(1)

The LDO class variable summarizes the information on the dependence between \( y_i \) and \( T_i \); missing data can be integrated out from equation (1)

\[ L_i(\cdot \mid T_i; \tau) = \prod_{i=1}^{n} \sum_{g=1}^{G} \sum_{s_i(\tau)} \left\{ \prod_{t=1}^{T_i} f_{y|s_s}(y_{it}^o \mid s_{it}, b_g; \tau) \delta_{s_1}^{(\tau)} \prod_{t=2}^{T_i} q_{s_{it-1}s_{it}(\tau)}(\tau) \right\} \pi_{ig}(T_i; \tau) \]

(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 $Time_{sero}$
Fixed and state-dependent parameters for the median

<table>
<thead>
<tr>
<th></th>
<th>lqHMM</th>
<th>lqHMM+LDO</th>
</tr>
</thead>
<tbody>
<tr>
<td># Par</td>
<td>36</td>
<td>33</td>
</tr>
<tr>
<td>Log-L</td>
<td>-1082.530</td>
<td>-1018.042</td>
</tr>
<tr>
<td>BIC</td>
<td>2377.802</td>
<td>2231.139</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>5.628 (5.074; 5.753)</td>
<td>6.043 (5.931; 6.114)</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>6.198 (6.014; 6.252)</td>
<td>6.416 (6.323; 6.502)</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>6.524 (6.393; 6.574)</td>
<td>6.719 (6.647; 6.825)</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>6.805 (6.719; 6.874)</td>
<td>7.040 (6.973; 7.215)</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>7.191 (7.084; 7.291)</td>
<td>-</td>
</tr>
<tr>
<td>Age</td>
<td>-0.003 (-0.007; 0.005)</td>
<td>0.004 (-0.001; 0.007)</td>
</tr>
<tr>
<td>Drugs</td>
<td>0.036 (-0.016; 0.110)</td>
<td>0.072 (-0.006; 0.145)</td>
</tr>
<tr>
<td>Packs</td>
<td>0.049 (0.014; 0.068)</td>
<td>0.042 (0.014; 0.054)</td>
</tr>
<tr>
<td>Partners</td>
<td>0.002 (-0.003; 0.012)</td>
<td>0.005 (0.000; 0.012)</td>
</tr>
<tr>
<td>CESD</td>
<td>-0.005 (-0.007; -0.001)</td>
<td>-0.004 (-0.006; -0.002)</td>
</tr>
<tr>
<td>$T_{\text{sero}}$</td>
<td>-0.110 (-0.126; -0.084)</td>
<td>-0.146 (-0.175; -0.119)</td>
</tr>
</tbody>
</table>

- 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 severe depression symptoms lead to decreasing CD4 counts
- CD4 counts decrease as the time since seroconversion increases
LDO class parameters

In the longitudinal data model

<table>
<thead>
<tr>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.497</td>
<td>-0.176</td>
<td>-0.070</td>
<td>0.033</td>
</tr>
<tr>
<td>(-0.667, -0.452)</td>
<td>(-0.200, -0.155)</td>
<td>(-0.098, -0.056)</td>
<td>(-0.023, 0.047)</td>
</tr>
</tbody>
</table>

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

In the LDO class model

<table>
<thead>
<tr>
<th>$\lambda_{01}$</th>
<th>$\lambda_{02}$</th>
<th>$\lambda_{03}$</th>
<th>$\lambda_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.062</td>
<td>1.113</td>
<td>4.089</td>
<td>-0.193</td>
</tr>
<tr>
<td>(-2.112, -0.241)</td>
<td>(0.013, 2.102)</td>
<td>(2.002, 5.299)</td>
<td>(-0.318, -0.065)</td>
</tr>
</tbody>
</table>

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


