Hidden Markov and mixture panel data models for ordinal variables derived from original continuous responses

FULVIA PENNONI
University of Milano-Bicocca
Department of Statistics and Quantitative Methods
Via Bicocca degli Arcimboldi 8, 20126 Milan
ITALY
fulvia.pennoni@unimib.it

GIORGIO VITTADINI
University of Milano-Bicocca
Department of Statistics and Quantitative Methods
Via Bicocca degli Arcimboldi 8, 20126 Milan
ITALY
giorgio.vittadini@unimib.it

Abstract: We evaluate the use of two different model formulations by proposing a modeling framework which extends the stochastic volatility models and the stochastic frontier models by considering an hidden Markov model formulation or a model made by a mixture of latent auto-regressive stochastic processes both of first order. Those models are suitable statistical tools to be fitted to many available panel data in various applicative cases. The proposed model formulation is especially tailored for ordinal data when they are derived as a grouping of a different scale. We show some features of the models estimation which is carried out by means of the maximum likelihood. In the illustrative example we recall the available function of the library LMest on the R environment which is tailored to carry out the estimation of the models. Further, provide some results of a case study to evaluate efficiency of a public organization by showing how the results can help policy makers.

Key–Words: Expectation-Maximization algorithm, global logits, generalized linear and mixed models, latent Markov model, R environment
2010 MSC: 91G30, 91B70, 62P30, 62M19

1 Introduction

In the following we show the use of two different kinds of latent variable models tailored for panel data. This type of use of such two models is relatively new in the literature of stochastic frontiers models. We describe how their formulation is especially appropriate when the interest lays on an ordinal response variable which categories are grouped of a finer scale. The latter is the case of an original continuous response which is suitable discretized. We recall some definition of the μ-th sample quantiles and that of the μ-th sample quantity quantile. We show that that the parameterization adopted for the distribution of each response variable based on global logits has many properties and interesting features in the context of phenomena evolving in time.

The paper is divided into two main parts. In the first we introduce the models and the notation and we state some relations related to the quantiles and quantity quantiles of a distribution. Then, we summarize some model features and illustrate some connections with the order statistics (David, 1970); secondly we summarize same details of the maximization procedures for both model formulations based on the log-likelihood. In the third section, we show some facilities of the available software to estimate the models in the R environment by considering a case study related to the estimation of technical efficiency. Then, we provide some conclusions.

2 Notation and specification of the proposed models

2.1 Background settings and notation

With reference to a sample of n units observed at T time occasions, let y_{it} be the ordinal response variable for unit i at occasion t. Let the number of categories denoted by J, and let $\boldsymbol{x}_{it}$ be a corresponding column vector of covariates, with $i = 1, \ldots, n$ and $t = 1, \ldots, T$. We denote by $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})$ the vector of response variables and by $\boldsymbol{X}_i = (\boldsymbol{x}_{i1} \cdots \boldsymbol{x}_{iT})$ the matrix of time-varying and time-constant covariates for unit i.

The model we formulate is based on the assumption that $y_{it} = G(y^*_{it})$, where y^*_{it} follows the model

$$y^*_{it} = \alpha_{it} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \eta_{it}, t = 1, \ldots, T,$$

for $i = 1, \ldots, n$, with η_{it} being independent error terms with a standard logistic or a standard normal distribution, and $G(\cdot)$ is a link function which models the relationship between each response variable y_{it}

ISBN: 978-1-61804-275-0

and the corresponding latent variable α_{it} and the vector of covariates $\boldsymbol{x}_{it}$. In such a case, it is a function of cut-points $\mu_1 \geq \cdots \geq \mu_{J-1}$ and it can be formulated as

$$G(y^*) = \begin{cases} 1 & y^* \leq -\mu_1, \\ 2 & -\mu_1 < y^* \leq -\mu_2, \\ \vdots & \vdots \\ J & y^* > -\mu_{J-1}. \end{cases}$$

The basic assumptions of the model are that for every sample unit i, $y_{it}^*, \ldots, y_{iT}^*$ are conditionally independent given $(\alpha_{i1}, \ldots, \alpha_{iT})$ and $\boldsymbol{X}_i$.

Due to the induced ordinal nature of the response variable we assume

$$\log \frac{p(y_{it} \geq j | \alpha_{it}, \boldsymbol{x}_{it})}{p(y_{it} < j | \alpha_{it}, \boldsymbol{x}_{it})} = \mu_j + \alpha_{it} + \boldsymbol{x}_{it}' \boldsymbol{\beta}, \qquad (1)$$

with $i = 1, \ldots, n$, $t = 1, \ldots, T$, $j = 2, \ldots, J$. These parameterization is based on global logits for the conditional distribution of each response variable and it is particularly suitable as we deal with an underling continuous outcome which is suitable discretized (McCullagh, 1980). Note that, the effect of the covariates and of the unobserved individual parameters do not depend on the specific response category.

This model is appropriate for those responses which are derived from an original continuous response variable at each time, for example when the interest lies in characterizing the distribution in terms of quantiles. In fact the μ-th quantile $\mu \in (0, 1)$ of a continuous random variable with density function f_y and distribution function F_y is defined as any number such that $F_y(\xi_\mu) = \mu$. The μ-th quantile function of Y supposing that f_y is strictly positive on the whole support of Y is defined as the $\mu_\theta = F_Y^{-1}(\mu)$ for $\mu \in (0, 1)$. Then, the μ quantile for the random variable Y is defined by a minimization of an appropriate loss function, see among others Koenker and Basset (1978). These authors show that the loss function can also be replaced by the empirical distribution function

$$F_n(y) = n^{-1} \sum_{i=1}^{n} I(y_i \leq y).$$

When the interest is considering the first-moment distribution that is

$$Q_Y(y) = \int_0^y \frac{t}{\mu} f_Y(t) dt$$

we are interested to the share of the total amount of the variable produced to the population with level of Y no greater than y. By considering the quantile function correspondent to the first incomplete moment defined as $Q_Y^{-1}(\mu)$ for $\mu \in (0, 1)$ we are led to what we call in the Italian literature quantity quantile for fixed μ, see among others Radaelli and Zenga (2008). They are widely used in the studies of income and wealth distribution, see among others Kleiber and Kotz (2003).

If we are dealing with a random sample of $Y_1, \ldots, Y_n$ and we consider the n sorted observations from Y the sample estimator is the sample quantity quantile defined as

$$\hat{\eta}_\mu = \inf\{y_{(i)} : \hat{Q}(y_{(i)}) \geq \mu\}$$

where

$$\hat{Q}(b) = \frac{\sum_{i: y_i \leq b} y_i}{T}$$

and T denotes the total amount. Therefore as showed by Radaelli and Zenga (2008) in order to obtain the μ-th sample quantity quantile the observations $y_{(i)}$, arrayed by increasing size, are summed until at least the share μ of the total is reached. It can be shown that the quantity quantiles of a given distribution F are the ordinary quantiles of another distribution G obtained from F by applying another suitably chosen function, for example the Lorenz function

$$L(p) = \frac{1}{E[Y]} \int_0^p F^{-1}(t) dt$$

for $p \in [0, 1]$.

2.2 Models details

The proposed approach extends the results on the quantity quantiles illustrated above to the cases of longitudinal data when repeated observations at different time occasions are available for the same subject. We conceptualize the model by including subject specific random effects and by modeling them according to two different types of formulations both relaying on the first order approximation of the underlying distribution.

In the following we illustrate the use two different types of distribution of the latent variable. The discrete latent process formulation assumes that, for all i, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iT})$ follows a first-order homogenous Markov chain with k states denoted by $\xi_1, \ldots, \xi_k$. This chain has initial probabilities π_h and transition probabilities $\pi_{h_1 h_2}$, with

$$\begin{aligned} \pi_h &= p(\alpha_{i1} = \xi_h), \quad h = 1, \ldots, k, \\ \pi_{h_1 h_2} &= p(\alpha_{i,t-1} = \xi_{h_1}, \alpha_{it} = \xi_{h_2}), \end{aligned}$$

where $h_1, h_2 = 1, \ldots, k$, $t = 2, \ldots, T$. It is assumed that every α_{it} is conditionally independent of $\alpha_{i1}, \ldots, \alpha_{i,t-2}$ given $\alpha_{i,t-1}$, but apart from this assumption, the distribution of $\boldsymbol{\alpha}_i$ is unconstrained. To

ensure identifiability we require that $\sum_h \pi_h = 1$ and $\sum_{h_2} \pi_{h_1 h_2} = 1$, $h_1 = 1, \ldots, k$ and one component of the support point is constrained to be zero. In such case a latent Markov (LM) model (Bartolucci, Farcomeni, Pennoni, 2013) with covariates results where the covariates affect the measurement model.

The above result is a generalization to the latent variable context of the very well known result that if we consider the order statistics $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ in a sample drawn from any continuous distribution, whose cumulative density function is strictly an increasing function of y, then, the random variables $Y_{(n)}, Y_{(n-1)}, \ldots, Y_{(1)}$ form a Markov chain so as the random variables $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$. Therefore, for a random sample of n continuous distributions the conditional distribution of $Y_{(s)}$ given $Y_{(r)}$ where $(r < s)$ is just the distribution of the $(s - r)$-th order statistic in a sample of $(n - r)$ drawn from the original distribution truncated at the left at $y = y_{(r)}$. A set of independences also holds when the original distribution is Gaussian as in the following.

The continuous latent process formulation assumes that the hidden response variables in $y^*_{i1}, \ldots, y^*_{iT}$ are conditionally independent given $\boldsymbol{X}_i$ and the latent process $\boldsymbol{\alpha}_i = (\alpha_{it}, \ldots, \alpha_{it})$. Another hypothesis is that every hidden variable and then every response variable, only depends on α_{it} and $\boldsymbol{x}_{it}$ and that the latent process $\boldsymbol{\alpha}_i$ has distribution given by a mixture of k AR(1) stochastic processes with common variance σ^2. According to the latter, we assume the existence of a discrete latent variable u_i, for $i = 1, \ldots, n$, having a distribution with k support points and mass probabilities $\pi_1, \ldots, \pi_k$ such that, when $u_i = h$ we assume that

$$\alpha_{i1} = \xi_h + \eta_{i1}, \quad i = 1, \ldots, n,$$

and that

$$\alpha_{it} = \xi_h + (\alpha_{i,t-1} - \xi_h)\rho_h + \eta_{it}\sqrt{1 - \rho_h^2},$$

where $i = 1, \ldots, n$, $t = 2, \ldots, T$, and $\eta_{it} \sim N(0, \sigma^2)$ for all i and t and (ξ_h, ρ_h) are parameters which for $h = 1, \ldots, k$ are estimated jointly with the common variance. To ensure identifiability of the model, we require that $\xi_1 = 0$ or, $\sum_h \xi_h \pi_h = 0$. We observe that when $h = 1$, the model is the latent auto-regressive model proposed by Chi and Reinsel (1989) and Heiss (2008), when $h > 1$ it is the mixture latent auto-regressive model proposed by Bartolucci, Bacci and Pennoni (2014). A summary of the parameters of the two proposed formulations is provided in Pennoni, Vittadini (2013).

It is important to mention that the choice between the two model formulations has important methodological implications and it depends on the problem of study. However, as showed later those two formulations may be easily compared in oder to choose the most appropriate model for the data. The proposed model framework can be useful in many practical applications to real data from different sources.

2.3 Main estimation features

An interesting feature of the proposal above is that likelihood-based estimation of the model parameter is feasible as for a sample of n independent units it is possible to consider the model log-likelihood of the form

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i | \boldsymbol{X}_i)$$

where $\boldsymbol{\theta}$ is the vector of all free parameters affecting the manifest distribution of the response vector $\boldsymbol{y}_i$ given all the observable covariates $\boldsymbol{X}_i$.

The maximum likelihood estimation of the model parameters in the case of a discrete formulation of the random process is performed by means of appropriately modify version of the Expectation-Maximization algorithm which first was implemented by Baum and Petrie in 1966 and also it is also known as Baum-Welch algorithm. It requires to maximize the conditional expected log-likelihood of the complete data which is formulated as the sum of three components representing the conditional distribution of the response variables given the covariates and the marginal distributions of the latent process. More details on the quantities involved in the estimation procedures may be found in Bartolucci, Farcomeni, Pennoni, (2013).

In the case of the continuous latent formulation of the random process a numerical optimization algorithm is used. The manifest distribution of the response variables given the covariates is expressed as a T-dimensional integral which is approximately computed by a quadrature method based on a series of q nodes properly chosen. Hence, in such a case, the computational burden increases according to the number of quadrature points chosen. For more details see Pennoni, Vittadini (2013) and Bartolucci, Bacci, Pennoni (2014). We also notice that the expression we compute for the manifest distribution of the response variables given the covariates based on q nodes is the same as that we get for the discrete random effect formulation based on q states. The manifest distribution is computed according to the recursions developed in the hidden Markov literature (Rabiner, 1990) and described in matrix notation by Bartolucci, Farcomeni, Pennoni (2013).

We provide also how to compute the standard errors for the estimated parameters by relaying on nu-

merical methods or on exact computational methods. One important result which exploits the observed information matrix is that of the missing information principle proposed by Tanner (1996) which allows for the estimated values with small extra code over that required by the maximum likelihood estimation. It is based on the results due to Orchard and Woodbury (1972) and Louis (1982) by which the observed information is equal to the complete information minus the missing information. Pennoni (2014) following Kivieri, Speed and Carlin (1984) suggests this approach by implementing it for the case of directed Gaussian acyclic graphical models with one hidden variable.

Another aspect of the proposal is the use of the goodness of fit indexes for model selection. In the case of a discrete formulation for the random parameters the Bayesian Information Criterion (Schwarz, 1978) has the advantage of considering the maximum log-likelihood of the model and to select a parsimonious model. It is based on the index

$$BIC = -2\ell(\hat{\boldsymbol{\theta}}) + g\log(n) \qquad (2)$$

where $\ell(\hat{\boldsymbol{\theta}})$ denotes the maximum log-likelihood of the model of interest and g is the number of parameters. Sometimes it may be not always the best choice but as stressed in the simulation study performed by Bacci, Pandolfi, Pennoni (2014) it performs properly in many simulated scenarios.

On the other hand, in the case of the mixture latent auto-regressive model the strategy for the choice of the number of components is more time consuming. For each component of the mixture the number of nodes have to be increased until the maximum of $\ell(\boldsymbol{\theta})$ does not significantly change with respect to the previous value obtained with q nodes. An optimal strategy is by increasing them by 10 and take the difference less than than 0.001. Then, once the optimal number of quadrature points are chosen, the model is estimated for an increasing number of states.

Another goal of the proposal is that the subject specific predictions may be gained for both models. For the case of the latent Markov model these are computed on the basis of the following expression:

$$\tilde{\alpha}_{it} = \sum_{h=1}^{k} = \hat{\pi}_h^* \hat{f}^{(h)}$$

for $i = 1, \ldots, n,\ t = 1, \ldots, T$ where $\hat{\pi}_h^*$ denotes the estimate of the stationary probability for the h latent state which depends on the transition probabilities $\hat{\pi}_{h_1,h_2}$ and $\hat{f}^{(h)}$ denotes the estimated posterior conditional distribution of the latent variables.

For the case of the mixture latent auto-regressive model these are computed as:

$$\tilde{\alpha}_{it} = \sum_{h=1}^{k} \sum_{m=1}^{q} (\widehat{w_{ih} z_{imt}})(\widehat{\xi}_h + \nu_m \hat{\sigma}),$$

for $i = 1, \ldots, n,\ t = 1, \ldots, T$, where $\widehat{w_{ih} z_{imt}}$ is the posterior density that subject i moves from state m_1 to state m_2 at occasion t given that $\mu_j = h$ and v_m denotes the m-th of the knots.

In such a way, it is possible to get some reliable estimates for each subject which are based on the observed covariates and on the selected clusters of the models. In fact, the proposed methodology allows also the graphical representation of such average predictions. This is an additional feature of the models which can be adopted also for policies or interventions in critical situations.

3 A case study and software implementation

The proposed models may be easily applied to real panel data under the statistical environment R (R Core Team, 2013). In the application proposed in Pennoni, Vittadini (2013) we consider a typical case which can be easily generalized to other settings requiring, for example, the evaluation of technical efficiency.

We consider the ratio between two dependent continuous variables of interest in the specific setting of the evaluation of the expenditures of public hospitals within a new reimbursement scheme adopted recently in one of the richest Italian region. In this framework, it common to consider two mainly measures of efficiency (see among others Hollingsworth, 2003 and Rosko and Mutter, 2007) which are the yearly number of discharges and the yearly revenues. The revenues are related to the number of discharges according to a perspective hospital's reimbursement scheme which has been recently introduced. The reimbursement is given on the basis of a tariff that the government sets at the beginning of each year which is related to groups of diagnosis. The rate received by the hospital for each admission depends on the patient's diagnosis. It is well stated in the literature that the system may give rises to some inappropriate behaviors made by the head of each yard or by the head of the hospital such that there is a negative trade-off between revenues and readmission and they have to be considered jointly to evaluate efficiency.

Therefore, in Pennoni, Vittadini (2013) we take into account the ratio between the yearly revenues and the yearly number of discharges for the data referred to the full population of patients for the general

ISBN: 978-1-61804-275-0

medicine ward. The latter is the one with the highest discharges and number of beds compared to the other wards in the region. We refer to this outcome as per capita revenue. We consider six available inputs (or covariates) which are time-varying: the number of beds of the hospital, the yearly hours of activity of the hospital's physicians, nurses, surgery rooms and of other employees which are not directly related with the surgery primary activity of the hospitals. In Table 1 and Table 2 we report the average values of each variables for every available year from 2008 to 2011 of the 110 hospitals considered in the analysis. It is interesting to note that the per capita revenues increase even if the average values of the input decreases over the years.

The use of the quantity quantiles allows us to investigate changes in the per capita revenues at interesting points of the distributions. Then we consider as response variable for the model in (1) the ordinal variable having four levels corresponding to the quantity quantiles of order 0.25, 0.5 and 0.75. The resulted categories of the derived ordinal variable are denoted as 'low', 'medium', 'high' and 'very high'. In Table 3 we show the empirical transition matrix of the response variables. Each row of this matrix shows the percentage frequencies of the four response categories at occasion t given the response at occasion $t-1$, with $t = 2, \ldots, T$.

Table 1: Average values of the outcome and of the input variables over the first two time occasions.

	Year	
Variable	2008	2009
pre capita revenue	2813.86	2886.57
beds (number)	45.51	45.39
physicans (hours)	245,597.44	247,104.11
nurses (hours)	481,504.42	485,475.39
others (hours)	460,843.22	459,612.36
surgey rooms (hours)	7,691.40	7,675.94

Table 2: Average values of the outcome and of the input variables over the last two time occasions.

	Year	
Variable	2010	2011
pre capita revenue	3011.58	3074.65
beds (number)	44.78	44.10
physicans (hours)	214,122.28	206,485.25
nurses (hours)	398,980.55	345,871.21
others (hours)	309,272.25	156,393.88
surgey rooms (hours)	8,144.78	7,940.05

Table 3: Conditional empirical distribution of the response variable at time t given the response at time $t-1$, with $t = 2, \ldots, T$ (percentage frequencies).

	Ratio at t			
Ratio at $t-1$	low	medium	high	very high
low	76.7	20.0	3.3	0.0
medium	20.7	48.3	24.1	6.9
high	3.4	31.0	44.8	20.7
very high	0.0	0.0	26.7	73.3

The proposed models may be estimated in a fast and easy way by a freely available package `LMest` (Bartolucci, Pandolfi, Pennoni 2014) in an improved version with the respect to previous one which is available from `http://cran.r-project.org/`. The choice of the `R` environment for data analysis and graphics provides a good degree of control and the user can also compare the models proposed above with other types of model proposed in the literature of stochastic frontier models such as those implemented on the `R` package called `Benchmaking`.

The main function which makes the applications of the model above very easy is called `est_lm_cov_manifest`. It requires to specify a matrix design for the response configurations according to the time occasions considered, the matrix of the input and the number of levels of the response variables which varies according to the chosen order of the quantity quantiles. Then, it requires to specify the number of states for which the model has to be fitted. The input `mod = 0` allows for the selection of the model with a discrete distribution for the random effects to be estimated. The input `mod = 1` allows for the model with a continuous distribution of the random effects. In the latter, the number of support points of the auto-regressive structure of the stochastic processes are specified by the input `q` to be set equal to an integer number. The other argument needed is the input `out se = TRUE` which allows to calculate the information matrix and the standard errors for the coefficients in model (1).

The main feature of the proposed approach compared with other standard models which measure technical efficiency is that the hospital can vary on the response variable because of the unobserved covariates such as general manager ability. The latter is a source of the so called unobserved heterogeneity which is important to take into account in many contexts of study. In this one, it is important as the head of the ward is the indeed lawfully responsible for all the activities performed in the hospital and we expect that his/her experience may influence the budgetary of the hospi-

tal.

As in other approaches we consider the translog function for the covariates (Christensen, Jorgenson, and Lau, 1973) in the estimation procedure. In general, when we are dealing with an ordinal response variable related to quantiles or to quantity quantiles as illustrate in Section 1 and we estimate the models with a discrete latent variable formulation as in Section 2.2 the optimal number of states is equal to the number of levels of the response variable. We estimate such model to the data at hand for an increasing number of k and each one is examined by the means of the BIC criterion. For the available data the model with $k = 4$ resulted in a BIC value equal to 819.891 when a log-likelihood has a maximum of -355.128 with 23 parameters.

Then, we specify the continuous latent variable formulation and we estimate the mixture latent auto-regressive model for an increasing number of quadrature points for each number of mixture components from 1 to 3. We select the number of quadrature points for the mixture latent auto-regressive model by means of the strategy illustrated in Section 2.3. We choose the following values: $q = 91$ for $k = 1$, $q = 81$ for $k = 2$ and $q = 111$ for $k = 3$. We consider the values of BIC for each value of k according to the chosen value of q. Then, the BIC leads to choose one mixture component that is the latent auto-regressive model. For the latter the log-likelihood at the maximum has value equal to -361.31 and the BIC is equal to 770.327 with 10 parameters.

In Table 4 we report the estimated parameters for both models with their standard errors. On the basis of the t-statistics that may be computed for the regression coefficients, we conclude that the first four covariates are significant on the latent Markov model with four latent states. On the other hand, while retaining the same sign only the first two covariates are significant under the mixture latent auto-regressive model with one component. The effect of the number of beds and of the working hours of physicians and nurses is positive, while the effect of working hours of the other staff of the hospital is negative, indicating that in the wards considered the main important features to explain efficiency are the first three. We conclude that the dimension of the hospital has a positive effect on the efficiency and that the hospital staff which is not directly related with the treatment of the patient may contribute to inefficiencies. The estimated initial probabilities relative to the chosen latent Markov model with $k = 4$ are the following: $\hat{\pi}_1 = 0.22$, $\hat{\pi}_2 = 0.31$, $\hat{\pi}_3 = 0.28$, $\hat{\pi}_4 = 0.19$. Under the mixture latent auto-regressive model with one component all the hospitals are in one latent class with high auto-correlation coefficient ($\hat{\rho}_1 = 0.911$) which is statistically significant (s.e. 0.361) and $\sigma^2 = 14.556$.

Table 4: Estimates of the parameters referred to the cut-points and the regression coefficients in equation (1), together with the corresponding standard errors for both types of models to which we refer as LM(4) and MLAR(1).

	LM(4)	MLAR(1)
$\hat{\mu}_1$	-47.073	-68.456
$\hat{\mu}_2$	-52.617	-76.136
$\hat{\mu}_3$	-58.841	-83.523
$\hat{\beta}_1$ beds	1.714	3.878
	(0.547)	(1.707)
$\hat{\beta}_2$ physicians	2.478	3.489
	(0.590)	(1.222)
$\hat{\beta}_3$ nurses	2.548	2.332
	(0.721)	(1.332)
$\hat{\beta}_4$ others	-1.259	-1.085
	(0.410)	(0.687)
$\hat{\beta}_5$ surgery rooms	0.137	0.552
	(0.283)	(0.637)

Due to the adopted parameterization the cut-points correspond to different levels of the propensity of the hospital to have high levels of efficiency. Therefore, we allow for different ways to detect efficiency changes. The value of the first cut-point is higher then the others, thus the first latent state corresponds to those hospitals with the highest propensity towards efficiency. The fourth latent state corresponds to those hospitals with the lowest propensity to be efficiency. Hence, they represents clusters of hospitals sharing the same propensity towards efficiency gains over the four year period considered.

Regarding the distribution of the latent process for the LM(4) model in Table 5 we report the estimates of the transition probability matrix.

Table 5: Estimates of the transition probabilities $\pi_{h_1 h_2}$ under the LM(4) model.

	$\hat{\pi}_{h_1 h_2}$			
h_2	$h_1 = 1$	$h_1 = 2$	$h_1 = 3$	$h_1 = 4$
1	0.911	0.042	0.047	0.000
2	0.064	0.936	0.000	0.000
3	0.000	0.037	0.907	0.056
4	0.000	0.000	0.089	0.911

Looking at the estimates of the parameters of the transition matrix we can see the evolution of the prob-

abilities of each state. In this way, we dispose of a characterization of the pathways of each of the four groups and we capture the behavior in a flexible manner.The matrix is not symmetric and the persistence in the same latent state for the entire period is high. The hospitals which have a medium/high level of efficiency i.e which are in latent state 2, in the previous year tend to become more efficient in the next year and those less efficient i.e. which are in latent state 4, in the previous year, tend to be more efficient as time goes.

Another feature of the models above is that by specify an additional input to the main function which estimate the models such as `output = TRUE` we get the most likely sequence for all sample units. Hence, it is also possible to dispose of a prediction of the individual effect for every hospital at each time occasion on the basis of the parameters estimates and on the available covariates (inputs). Once they are depicted, for example, it is interesting to note those predicted profiles trajectories which are less regular then others so that it is possible to quickly identify those hospitals with suddenly changes and to correct in advance some opportunistic behaviors of the hospitals in a cost effective strategy. By the inspection of the graph of the predicted values we can notice if the single predicted profile trajectories are less regular under one model compared to the other. Thus, this means that we can detect in a more appropriate way with this model compared to the other the changes observed in the hospitals which are due to events which are not observed through the covariates. Moreover, according to them, it is also possible to rank the structures from the best to the worst performer in terms of potentially efficiency gains.

4 Conclusion

In this paper, we have proposed the use of two special kinds of latent variable models for analyzing longitudinal data when the ordinal response variables are derived as a grouping of a different scale. First, we show that the model formulation we derive is suitable when the responses are derived from an original continuos response variable such as when the interest lies in characterizing the distribution in terms of quantiles. Then, we show the two model formulations take into account that in this way we are dealing with order statistics. The first one assuming a discrete distribution gives rise to a latent Markov model which is not very complex to fit. It is more natural in many contexts and very suitable for classification even if the number of parameters increases with the number of latent states. The second model formulation relies on a continuous distribution for the unobserved heterogeneity. It gives rise to a mixture latent auto-regressive model which is more complex to fit. Maximum likelihood estimation of the model parameters is performed by a joint use of the Expectation-Maximization algorithm and of the Newton-Raphson algorithm. Standard errors for the parameter estimates are also obtained. The number of latent states are selected by considering the BIC index for the latent Markov model and by an appropriate strategy for the mixture components. We present a way to derive estimated predictions of the individual effects for every unit at each time occasion on the basis of the parameter estimates. In the last section the models are applied to real panel data to investigate technical inefficiencies of public hospitals of one of the richest Italian region. The response variable of interest is obtained by considering the revenues and the number of discharges which are relevant in the reimbursement scheme adopted by the government of the region.

By the use of this case study we also illustrate one of the main function of the R package `LMest` which provides the facilities to estimate the proposed models. We show the flexibility of the models and how they can be useful to monitor the evolution of the performance of the selected clusters of hospitals over time which share the same propensity towards efficiency gains. The results gained by the application of the proposed models may be used to support decision makers to improve or to intervene on the process under study.

Acknowledgements: We acknowledge “Finite mixture and latent variable models for causal inference and analysis of socio-economic data” (FIRB - Futuro in ricerca) funded by the Italian Government (RBFR12SHVV). F. Pennoni also thanks the financial support of the STAR project “Statistical models for human perception and evaluation”, University of Naples Federico II.

References:

[1] Aigner, D., Lovell, C. A. K. and Schmidt, P. (1977), Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics*, **36**, 21-37.

[2] Bacci, S., Pandolfi, S., and Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, **8**, 125-145.

[3] Bartolucci, F., Farcomeni A., Pennoni, F. (2014). Latent Markov models: a review of a general

framework for the analysis of longitudinal data with covariates, *with discussion*, *Test*, **23**, 433-465.

[4] Bartolucci F, Pandolfi S., Pennoni F. (2014). Fit Latent Markov models in basic versions. R package version 2.0. `http://CRAN.R-project.org/package=LMest`.

[5] Bartolucci, F., Bacci, S., Pennoni, F. (2014). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models, *Journal of the Royal Statistical Society: Series C*, **63**, 267-288.

[6] Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*, Chapman and Hall/CRC press, Boca Raton.

[7] Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41, 164 -171.

[8] Chi, E. and Reinsel, G. (1989). Models for longitudinal data with random effects and AR(1) errors, *Journal of the American Statistical Association*, **84**, 452-459.

[9] Christensen, L., Jorgenson, D., Lau, L. (1973). Transcendental logarithmic production frontiers, *Review of Economics and Statistics*, **55**, 28–45.

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

[11] David H. A. (1970). *Order Statistics*, Wiley, New York.

[12] Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for microeconometric panel data, *Journal of Applied Econometrics*, **23**, 373-389.

[13] Herwartz, H., Strumann, C. (2014). Hospital efficiency under prospective reimbursement schemes: an empirical assessment for the case of Germany, *European Journal of Health Economics*, **15**, 175-186.

[14] Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care, *Health care management science*, **6**, 203-218.

[15] Green, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model, *Journal of Econometrics*, **126**, 269–303.

[16] Greene, W. (2009). The econometric approach to efficiency analysis, in H. O. Fried, C. A. K. Lovell and S. S. Schmidt (Eds), *The measurement of productive efficiency techniques and applications*, Oxford University Press: Oxford, 92-251.

[17] Kiiveri H. T., Speed T. P., Carlin J. B. (1984). Recursive causal models. *J. Austral. Math. Soc. Ser. A*, **36**, 30-52.

[18] Kleiber C., Kotz S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley Series in Probability and Statistics. Wiley, New York.

[19] Koenker R, Bassett G. J. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

[20] Kumbhakar, S. C., Lien, G. and Brian J. (2014). Technical efficiency in competing panel data models: a study of Norwegian grain farming, *Journal of Productivity Analysis*, **41**, 321-337.

[21] Louis T. A. (1982). Finding the observed information matrix when using the EM-algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226-233.

[22] McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109–142.

[23] Pennoni, F. (2014). *Issues on the estimation of latent variable and latent class models*. Scholars' Press, Saarbucken.

[24] Pennoni, F., Vittadini, G. (2013). Two competing models for ordinal longitudinal data with time-varying latent effects: an application to evaluate hospital efficiency. *QdS, Journal of methodological and applied statistics*, **15**, 53-68.

[25] Orchard T., Woodbury M. A. (1972). A Missing Information Principle: Theory and Applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 697-715.

[26] R Development Core Team (2013). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing.

[27] Rabiner, L. (1990). A tutorial on Hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, **53**, 267-296.

[28] Radaelli, P., Zenga, M (2008). Quantity quantiles linear regression, *Statistical methods and applications*, **17**, 455-469.

[29] Rosko, M.D., Mutter, R. L. (2007). Stochastic frontier analysis of hospital inefficiency: a review of empirical issues and an assessment of robustness, *Medical care research and review*, **65**, 131–166.

[30] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.

[31] Tanner M.A. (1996). *Tools for statistical inference*. New York: Springer.

[32] Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for time series: an introduction using R*. Springer-Verlag, New York.