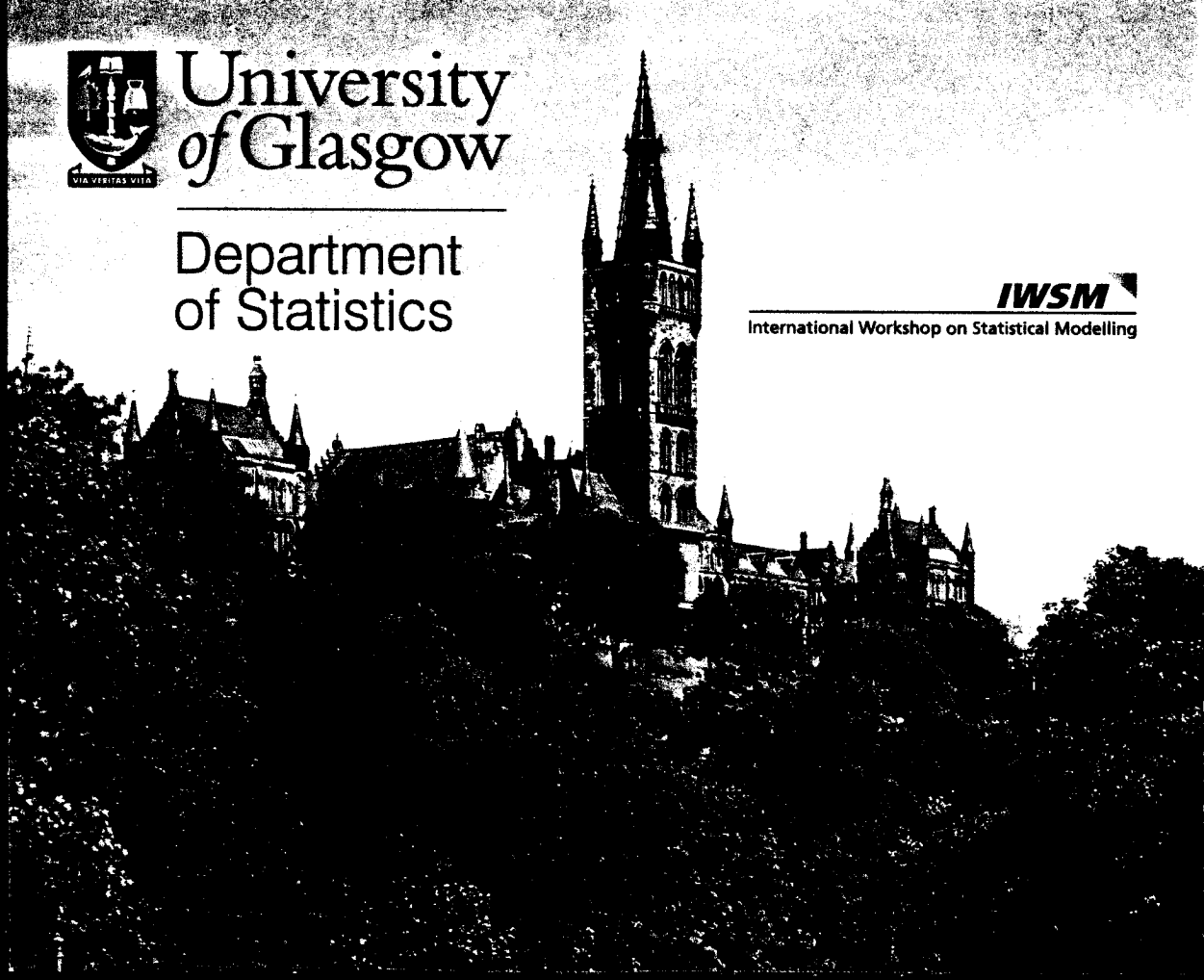

University of Glasgow

Department of Statistics

IWSM

International Workshop on Statistical Modelling

25th International Workshop on Statistical Modelling

University of Glasgow 5-9th July, 2010

Proceedings

Proceedings of the 25th International Workshop on Statistical Modelling

July 5-9, 2010

Glasgow

Adrian W. Bowman
(editor)

Preface

In 2010, the International Workshop on Statistical Modelling celebrates its 25th birthday. Twenty-five years old is a good age to be, as it combines maturity with vitality, and both of these attributes are evident in abundance in this year's workshop. The invited papers demonstrate both the current exciting developments in methodology and the key role which statistical modelling plays in a very wide variety of modern applications. This is amplified in both breadth and depth in the large number of contributed papers at this year's workshop. Some aspects of the programme honour the past and celebrate the origins and development of statistics, as befits a birthday event. However, very importantly, there is a major focus on the future, marked in particular by a substantial number of student contributions. It is good to see that statistical modelling is in very good health.

This year the IWSM comes to Glasgow, a city with a rich history and a dynamic culture. We hope that you will take full advantage of your stay here to enjoy the character of the city and a variety of social events on the workshop programme aim to help you in doing this. Glasgow is also located close to some spectacular Scottish scenery and we hope that you will have time to sample a little of this.

However, Glasgow also has a rich history in statistics and the Department of Statistics in the University of Glasgow is particularly pleased to be able to host the IWSM. The aims of the workshop coincide with the aims of the Department, to promote the subject of statistics and its application to important scientific problems, and so we look forward very much to meeting you and interacting with you.

The scientific success of the workshop depends on the participants, although this is focussed by the scientific committee whose contributions are very much appreciated. The organisational success of the workshop necessarily depends on a much smaller number of people, and the pivotal role is played by our local organiser, Claire Ferguson. With the very able assistance of Sarah Barry, plus a variety of administrative and IT staff, Claire has put very considerable time, energy and expertise into the preparations for the workshop. We hope that you enjoy the fruits of these labours. For my own part, I would also like to thank Ludger Evers for invaluable assistance in the technicalities of preparing the conference proceedings.

So welcome to Glasgow, and enjoy the workshop!

Adrian Bowman
Glasgow, July 2010

Markov-switching autoregressive latent variable models for longitudinal data

Silvia Bacci[1], Francesco Bartolucci[2], Fulvia Pennoni[3]

[1] Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy, silvia.bacci@stat.unipg.it
[2] Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy, bart@stat.unipg.it
[3] Department of Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy, fulvia.pennoni@unimib.it

Abstract: We propose a generalization of the autoregressive latent variable models for longitudinal data based on an AR(1) process to represent the effect of unobservable factors on the response variables. The generalization is based on assuming that the latent process follows a Markov-switching AR(1) process with correlation coefficient depending on the regime of the chain. Some particular cases are discussed in detail and illustrated by an application to a longitudinal dataset about self-evaluation of the health status.

Keywords: Latent Markov model; Ordinal variables; Numerical integration.

1 Introduction

In the analysis of longitudinal data, an important aspect to be taken into account is how to represent the effect that unobservable factors have on the occasion-specific response variables in addition to the effect of observable covariates. The simplest approach is based on the inclusion, in the model of interest, of individual-specific random intercepts. In this way, however, the effect of unobservable factors is assumed to be time constant. A natural way to relax this assumption is by assuming that, for each subject, there are occasion-specific random effects which follow an AR(1) process (Chi and Reinsel, 1989); the resulting model will be referred to as latent autoregressive (LAR) model. An alternative formulation is based on the inclusion of a sequence of discrete latent variables which follow a first-order Markov chain. In this way, a Latent Markov (LM) model (Wiggins, 1973) with covariate results. For a review on LM models see Bartolucci *et al.* (2010) and for an instance of a complex model formulated following this approach see Bartolucci and Farcomeni (2009).
The main advantage of the LAR formulation is that it retains a parsimony close to that of the corresponding random effect model. Moreover, in certain applications it is natural to represent the error terms by continuous

rather than discrete random variables. On the other hand, estimating the resulting model may be problematic from the computational point of view (Heiss, 2008). The model based on the LM formulation naturally provides a classification of subjects into a reduced number of groups, is easier to estimate, and may reach a better fit. However, this model is usually less parsimonious. It is also worth noting that a Markov chain is able to approximate adequately a continuous process and then the model based on the LM formulation may be seen as a semi-parametric version of the model based on the AR(1) process. The issue of the comparison between the two approaches above is related to that of the comparison between a standard random effect model and its latent class version in contexts simpler to the present one; see Lindsay *et al.* (1991) and Greene and Hensher (2003).
In this paper, we formulate a model for longitudinal data which is based on the assumption that the error terms follow a Markov-switching AR(1) process (Hamilton, 1989). In particular, we assume that a set of different regimes are possible, with each regime corresponding to a different value of the correlation coefficient. How a subject moves between regimes is governed by an unobservable Markov chain which is time-homogenous. In this way, we extend the LAR model by allowing the correlation coefficient to be different between subjects and occasions. Moreover, we expect that the resulting model has a fit comparable to that of a model based on a LM formulation, but it is more parsimonious. Two versions of the proposed model are discussed in detail. In the former, the autoregressive correlation coefficient may be different between subjects, but not between occasions. In the second version, instead, each subject randomly moves between different regimes. Both versions are estimated by the maximum likelihood method, which is implemented on the basis of an algorithm similar to the sequential numerical integration algorithm proposed by Heiss (2008).
The paper is organized as follows. In the following section we introduce the basic notation and describe the LAR formulation for longitudinal data. In Section 3 we outline the proposed extension, whereas the results of an illustrative application based on a dataset about self-evaluation of the health status are briefly illustrated in Section 4.

2 Preliminaries

Let y_{it} be the response variable observed at occasion $t = 1, \ldots, T$ for subject $i = 1, \ldots, n$ and let $\boldsymbol{x_{it}}$ be a corresponding vector of covariates.
The model based on the LAR formulation for these variables assumes that, for every subject i, $y_{i1}, \ldots, y_{iT}$ are conditionally independent given the covariates $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}$ and a sequence of latent variables $u_{i1}, \ldots, u_{iT}$ which follows an AR(1) process. In particular, we assume that $u_{i1} \sim N(0, \sigma^2)$ and

that, for $t > 1$,

$$u_{it}|u_{i,t-1} = \rho u_{i,t-1} + \varepsilon_{it}, \quad (1)$$

$$\varepsilon_{it} \sim N[0, \sigma^2/\sqrt{1-\rho^2}]. \quad (2)$$

An important point is how to model the conditional distribution of each response variable y_{it} given u_{it} and $\boldsymbol{x_{it}}$. For instance, in the case of ordinal responses with l categories, that we will consider in our application, a natural parametrization is based on cumulative (or global) logits:

$$\log \frac{p(y_{it} > j|u_{it}, \boldsymbol{x}_{it})}{p(y_{it} \leq j|u_{it}, \boldsymbol{x}_{it})} = \mu_j + u_{it} + \boldsymbol{x}_{it}'\boldsymbol{\beta}, \quad j = 1, \ldots, l-1.$$

The main difference with the LM formulation is that in the latter the latent process follows a Markov chain with k states, with the following parameters: $k-1$ support points (the first is fixed at 0), in addition to $k-1$ initial probabilities, and $k(k-1)$ transition probabilities. The LAR model uses, instead, only 2 parameters for the latent process.

3 The proposed model

We generalize the LAR model presented in the previous section in order to allow for a different correlation between time occasions and subjects. For this aim we exploit the general framework of Markov-switching autoregressive models (Hamilton, 1989), where the correlation coefficient depends on an unobserved Markov chain.

3.1 Model assumptions

The proposed model (named SW-LAR) is formulated as in Section 2, with assumptions (1) and (2) substituted by

$$u_{it}|u_{i,t-1}, v_{it} = \rho_{v_{it}} u_{i,t-1} + \varepsilon_{it},$$

$$\varepsilon_{it}|v_{it} \sim N[0, \sigma^2/\sqrt{1-\rho_{v_{it}}^2}],$$

where the latent process $v_{i1}, \ldots, v_{iT}$ follows a Markov-chain with k latent states corresponding to the correlation coefficients $\rho_1, \ldots, \rho_k$. This process is characterized by the vector of initial probabilities $\boldsymbol{\lambda}$, with elements λ_v, $v = 1, \ldots, k$, and the transition probability matrix $\boldsymbol{\Pi}$, with elements $\pi_{v_0 v}$, $v_0, v = 1, \ldots, k$. Note that every latent variable u_{it} has marginal distribution $N(0, \sigma^2)$ as in the LAR model.
It is worth noting that by imposing constraints on k, $\boldsymbol{\lambda}$, or $\boldsymbol{\Pi}$, special cases of the SW-LAR model result. In particular:

1. with $k = 1$ the LAR model described in Section 2 is obtained. The correlation coefficient is then the same for all subjects and occasions.

TABLE 1. Results from the fitting of models LAR, SW-LAR$_1$, and SW-LAR$_2$.

	LAR	SW-LAR$_1$	SW-LAR$_2$
μ_1	7.3270	9.1515	7.6452
μ_2	4.1949	5.2750	4.3014
μ_3	1.0229	1.2479	0.9076
μ_4	-2.3763	-3.0282	-2.6919
female	-0.0572	0.0440	-0.0591
non white	-1.8515	-2.2072	-1.8758
education	1.5882	1.9401	1.6746
age	-0.1012	-0.1207	-0.0929
σ	2.9159	3.9973	3.2414
ρ_1	0.9550	0.4889	0.4414
ρ_2	–	0.9758	1.0000
λ_1	1.0000	0.2406	0.1268
λ_2	–	0.7594	0.8732
log-likelihood	-8884.7	-8795.6	-8818.2
# parameters	10	12	12
BIC	17838	17674	17719

In order to maximize $\ell(\boldsymbol{\theta})$ we implemented a numerical algorithm which, for the moment, may be only used to deal with LAR, SW-LAR$_1$, and SW-LAR$_2$ models. Future research will be devoted to the implementation of an algorithm to estimate the more general SW-LAR model and to obtain standard errors for the parameter estimates.

4 Application

The data used for the illustrative application come from the Health and Retirement Study conducted by the University of Michigan (for a detailed description see `http://www.rand.org/labor/aging/dataprod`). In particular, we considered a set of 1000 American people who self-evaluated their health status over 8 occasions. The health status is measured on a scale based on five grades: poor, fair, good, very good, and excellent. For each subject, some covariates are available: *gender*, *race*, *education*, and *age at each occasion of interview*.

We fitted three different models on these data: LAR, SW-LAR$_1$ ($\boldsymbol{\Pi} = \boldsymbol{I}$) with $k = 2$ latent states, and SW-LAR$_2$ ($\boldsymbol{\Pi} = \mathbf{1} \otimes \boldsymbol{\lambda}'$) with the same number of latent states. The main results are given in Table 1.

We note that the SW-LAR$_1$ model has the highest log-likelihood and the smallest value of the BIC index (Schwarz, 1978). With the inclusion of only two more parameters, this model shows a much better fit than the LAR model. The estimates of the two correlation coefficients under this model

2. If the transition matrix is equal to an identity matrix, i.e. $\mathbf{\Pi} = \boldsymbol{I}$, the SW-LAR$_1$ model is obtained. Under this model, the correlation coefficient may be different between subjects belonging to different latent states, but not between occasions.

3. If the transition matrix has constant rows containing the initial probabilities, i.e. $\mathbf{\Pi} = \mathbf{1} \otimes \boldsymbol{\lambda}'$, the SW-LAR$_2$ model results, under which the correlation coefficient may change between subjects and occasions, since each subject randomly moves between different regimes.

3.2 Model estimation

We estimate the model parameters by maximizing the corresponding log-likelihood which is given by

$$\ell(\boldsymbol{\theta}) = \sum_i \log p(\boldsymbol{y}_i|\boldsymbol{X}_i),$$

where $\boldsymbol{\theta}$ is a short-hand notation for all model parameters, $\boldsymbol{y}_i$ is the response vector with elements y_{it}, $t = 1, \ldots, T$, and $\boldsymbol{X}_i$ is the corresponding matrix of covariates made of the vectors $\boldsymbol{x}_{it}$.
A crucial point is how to compute the *manifest probability* or *density* $p(\boldsymbol{y}_i|\boldsymbol{X}_i)$, which is based on a T-dimensional integral. For this aim we implemented an algorithm which is related to the sequential numerical integration method of Heiss (2008).
Let $q_{it}(u, v) = p(u_{it} = u, v_{it} = v, y_{i1}, \ldots, y_{it})$ and note that, for $t > 1$, this probability may be expressed as

$$q_{it}(u, v) = f(y_{it}|u) \sum_{v_0} \pi_{v_0 v} \int_{\Re} q_{i,t-1}(u_0, v_0) g(u|u_0, v) du_0,$$

with

$$q_{i1}(u, v) = p(y_{i1}|u) \rho_v g(u),$$

where $f(y_{it}|u) = p(y_{it}|u_{it} = u)$, $g(u)$ denotes the density function for the distribution of u_{i1}, and $g(u|u_0, v)$ denotes that for the distribution of u_{it} given $u_{i,t-1} = u_0$ and $v_{it} = v$, with $t > 1$. The algorithm we implemented is based on computing first $q_{i1}(u, v)$ and then $q_{it}(u, v)$ for $t = 2, \ldots, T$ by a suitable Gaussian quadrature. These probabilities are computed for u equal to every node of the quadrature and $v = 1, \ldots, k$. At the end, we obtain

$$p(\boldsymbol{y}_i|\boldsymbol{X}_i) = \sum_v \int_{\Re} q_{iT}(u, v) du,$$

again computed by a suitable quadrature. Note that, this algorithm closely resembles the recursive algorithm commonly used for the maximum likelihood estimation of hidden Markov models (Baum *et al.*, 1970).

TABLE 1. Results from the fitting of models LAR, SW-LAR$_1$, and SW-LAR$_2$.

	LAR	SW-LAR$_1$	SW-LAR$_2$
μ_1	7.3270	9.1515	7.6452
μ_2	4.1949	5.2750	4.3014
μ_3	1.0229	1.2479	0.9076
μ_4	-2.3763	-3.0282	-2.6919
female	-0.0572	0.0440	-0.0591
non white	-1.8515	-2.2072	-1.8758
education	1.5882	1.9401	1.6746
age	-0.1012	-0.1207	-0.0929
σ	2.9159	3.9973	3.2414
ρ_1	0.9550	0.4889	0.4414
ρ_2	-	0.9758	1.0000
λ_1	1.0000	0.2406	0.1268
λ_2	-	0.7594	0.8732
log-likelihood	-8884.7	-8795.6	-8818.2
# parameters	10	12	12
BIC	17838	17674	17719

In order to maximize $\ell(\boldsymbol{\theta})$ we implemented a numerical algorithm which, for the moment, may be only used to deal with LAR, SW-LAR$_1$, and SW-LAR$_2$ models. Future research will be devoted to the implementation of an algorithm to estimate the more general SW-LAR model and to obtain standard errors for the parameter estimates.

4 Application

The data used for the illustrative application come from the Health and Retirement Study conducted by the University of Michigan (for a detailed description see `http://www.rand.org/labor/aging/dataprod`). In particular, we considered a set of 1000 American people who self-evaluated their health status over 8 occasions. The health status is measured on a scale based on five grades: poor, fair, good, very good, and excellent. For each subject, some covariates are available: *gender*, *race*, *education*, and *age at each occasion of interview*.

We fitted three different models on these data: LAR, SW-LAR$_1$ ($\boldsymbol{\Pi} = \boldsymbol{I}$) with $k = 2$ latent states, and SW-LAR$_2$ ($\boldsymbol{\Pi} = \mathbf{1} \otimes \boldsymbol{\lambda}'$) with the same number of latent states. The main results are given in Table 1.

We note that the SW-LAR$_1$ model has the highest log-likelihood and the smallest value of the BIC index (Schwarz, 1978). With the inclusion of only two more parameters, this model shows a much better fit than the LAR model. The estimates of the two correlation coefficients under this model

are rather different; in particular we have an estimate equal to 0.49 for the 24% of subjects and equal to 0.98 for the remaining 76% of subjects. These two different levels of correlation correspond to two different levels of persistence of the effect of unobservable factors on the response variables.

Acknowledgments

F. Bartolucci and F. Pennoni acknowledge the financial support from the "Einaudi Institute for Economics and Finance" (Rome - IT) and from PRIN 2007.

References

Bartolucci, F., and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, **104**, 816-831.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2010). An overview of latent Markov models for longitudinal categorical data. *Technical report* `http://arxiv.org/abs/1003.2804`.

Baum, L.E., 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.

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

Greene, W.H., and Hensher, D.A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research, Part B*, **37**, 681-698.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357-384.

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

Lindsay, B., Clogg, C.C., and Greco, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, **86**, 96-107.

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

Wiggins, L.M. (1973). *Panel Analysis: Latent probability models for attitude and behaviour processes*. Amsterdam: Elsevier.