Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 42: 1127–1135, 2015
doi: 10.1111/sjos.12155

Generalized Linear Mixed Models Based on Latent Markov Heterogeneity Structures

ALESSIO FARCOMENI

Department of Public Health and Infectious Diseases, Sapienza - University of Rome

ABSTRACT. We describe a generalized linear mixed model in which all random effects may evolve over time. Random effects have a discrete support and follow a first-order Markov chain. Constraints control the size of the parameter space and possibly yield blocks of time-constant random effects. We illustrate with an application to the relationship between health education and depression in a panel of adolescents, where the random effects are highly dimensional and separately evolve over time.

Key words: hidden Markov model, longitudinal data analysis, mixed models, time-varying random effects, unobserved heterogeneity

1. Introduction

We propose a generalized linear mixed model (GLMM) with multidimensional random effects, all separately time-varying. Each random effect has got a discrete distribution, with a specific number of support points. They evolve over time according to arbitrarily dependent first-order Markov chains. A block random model groups random effects driven by the same latent process. A natural grouping exists, for instance, when there are cluster indicators. A drawback of the general model is that the number of latent states is potentially large with high-dimensional random effects. The number of parameters will be controlled with appropriate constraints for the cluster-specific intercepts or probability masses. A sample code is available from the author upon request, and an optimized version will be included in `R` library `LMest` (Bartolucci *et al.*, 2015). For background on GLMMs, see for instance McCulloch & Searle (2001) & Diggle *et al.* (2002). There are few GLMM formulations where flexible time-varying structures are allowed for the random effects. Discrete supports are used in Altman (2007) and Maruotti & Rydén (2009), but only a single univariate latent chain is time-varying in these works. Bartolucci & Lupparelli (2015) specify two levels of nested time-varying random effects. In our opinion, discrete random effects have three advantages over continuous ones (Bartolucci *et al.*, 2014): first, estimates are easily interpretable and provide clusters of subjects. Second, marginals can be obtained *exactly*, as integrals reduce to sums. Finally, they can often approximate continuous distributions without having to specify parametric forms. Limitations include that continuous random effects may be more easily justified (especially with continuous outcomes), and that inferential procedures might yield local optima. This problem is usually successfully tackled via multistart. The rest of the paper is as follows: in the next section, we set up our GLMM with random effects following a first-order Markov chain, and restricted versions which are also interpretable. In Section 3, we discuss inferential issues. A simulation study in Section 4 illustrates and compares the approach. In Section 5, we report a real data analysis and give concluding remarks in Section 6.

2. Model formulation

Suppose n subjects have been repeatedly observed at T occasions. Let y_{it}, $i = 1, \ldots, n$; $t = 1, \ldots, T$ denote an outcome of interest, arising from a general exponential family with usual notation. Our GLMM assumes therefore that

$$\begin{cases} f(y_{it}|\boldsymbol{\alpha}_{it}, \boldsymbol{\beta}, \psi) = \exp\{(y_{it}\eta_{it} - c(\eta_{it}))/(a(\psi) - b(y_{it}, \psi))\} \\ g(\eta_{it}) = \boldsymbol{z}'_{it}\boldsymbol{\alpha}_{it} + \boldsymbol{x}'_{it}\boldsymbol{\beta}; \end{cases} \tag{1}$$

where $\boldsymbol{x}_{it}$ is a vector of d dimensional covariates associated with fixed effects $\boldsymbol{\beta}$, and $\boldsymbol{z}_{it}$ is a vector of covariates associated with random effects $\boldsymbol{\alpha}_{it}$. The use of a general link function $g(\cdot)$ lets us specify non-canonical link functions when desired. We assume there are R blocks of random effects for some $R \geq 1$, that $\boldsymbol{z}_{it}$ and $\boldsymbol{\alpha}_{it}$ are partitioned accordingly, and that block r is q_r dimensional, $r = 1, \ldots, R$. The variables within each block are driven by the same random process. Consequently, a *vector* of q_r support points is associated with each state of the r-th time-varying latent process. If $q_r > 1$, each dimension takes possibly different values, but all dimensions are indicized by the same latent process. More formally, let α_{itrh} denote the subject and time-specific random effect associated with the h-th covariate in the r-th block. We assume that, for $r = 1, \ldots, R$ and $h = 1, \ldots, q_r$, α_{itrh} is a discrete random variable with k_r support points, denoted with $\xi_{1rh}, \ldots, \xi_{k_r rh}$, and collected in the vectors $\boldsymbol{\xi}_{1r}, \ldots, \boldsymbol{\xi}_{k_r r}$. It is finally assumed that random effects evolve over time according to a first-order homogeneous Markov chain. We specify an initial joint distribution

$$\Pr\left(\bigcap_{r=1}^{R}\bigcap_{h=1}^{q_r} \alpha_{i1rh} = \xi_{c_r rh}\right) = \Pr(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_{\boldsymbol{c}}) = \pi_{c_1,\ldots,c_R}, \tag{2}$$

for $c_r = 1, \ldots, k_r$, $r = 1, \ldots, R$. These parameters sum to the unity and are collected in a vector $\boldsymbol{\pi}$. The transition matrix $\boldsymbol{\Pi}$ is defined analogously as, for $t > 1$, it is assumed that

$$\Pr(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\boldsymbol{d}} \,|\, \boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\xi}_{\boldsymbol{c}}) = \pi_{\boldsymbol{cd}}, \tag{3}$$

where $\boldsymbol{c} = \{c_1, \ldots, c_R\}$, similarly for $\boldsymbol{d}$, and $\sum_{\boldsymbol{d}} \pi_{\boldsymbol{cd}} = 1$ for all $\boldsymbol{c}$. The parameters $\pi_{\boldsymbol{cd}}$ are collected in the transition matrix $\boldsymbol{\Pi}$, where each possible value of $\boldsymbol{c}$ yields a row of the matrix. The joint distribution at time t is obtained as

$$\Pr(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\boldsymbol{c}}) = \left(\boldsymbol{\pi}'\boldsymbol{\Pi}^{t-1}\right)_{c_1,\ldots,c_R}. \tag{4}$$

The marginal distribution of α_{itrh}, which has support $\xi_{1rh}, \ldots, \xi_{k_r rh}$, is obtained with opportune sums of the expression earlier. To summarize, we have $\sum_{r=1}^{R} q_r k_r$ free parameters associated with latent intercepts and d parameters $\boldsymbol{\beta}$ associated with fixed effects. Furthermore, $\prod_{r=1}^{R} k_r - 1$ free parameters are used to model the initial distribution and $\prod_{r=1}^{R} k_r \prod_{r=1}^{R}(k_r - 1)$ for the transition matrix.

In many applications, the dimensionality of $\boldsymbol{Z}$ is reasonable, and each random effect is based on a low number of support points. The number of free parameters is anyway bound to explode as the dimensionality of the random effects distribution or the number of support points are increased. Fine tuning is desirable to balance between complexity and goodness-of-fit. Restrictions can be put on the latent distribution to this end. We outline here assumptions that have a direct meaning in terms of model interpretation. One possibility is to assume that some random effects share the same support points. This is meaningful for random effects associated to continuous variables, or belonging to different groups with the same number of latent states. Another possibility is to assume that the support points within the r-th block

are k_r (rather than $q_r k_r$), so that for $h = 2, \ldots, q_r$ $\boldsymbol{\xi}_{rh} = \sigma_{rh}(\boldsymbol{\xi}_{r1})$, where $\sigma_{rh}(\cdot)$ denotes a known or unknown permutation. These constraints are particularly useful for large groups of binary indicators. Constraints can also involve the probability masses. A natural possibility is to assume (or at least test) independence, or block independence, of the random effects. Under the assumption of independence, the joint distribution of the random effects is obtained as the product of the marginal distributions. Formally

$$\Pr(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\boldsymbol{c}}) = \prod_{r=1}^{R} \prod_{h=1}^{q_r} \Pr\left(\alpha_{itrh} = \xi_{c_r rh}\right). \tag{5}$$

This assumption implies that the initial distribution is parameterized as

$$\Pr(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_{\boldsymbol{c}}) = \prod_{r=1}^{R} \pi_{c_r r}, \tag{6}$$

with obvious notation for the marginals $\pi_{c_r r} = \Pr(\alpha_{i1rh} = \xi_{c_r rh})$, $r = 1, \ldots, R$, collected in $\boldsymbol{\pi}_r$. An assumption along the lines of (6) is used for the transition matrix. This automatically guarantees (5) for $t \geq 1$. The number of parameters used to model the latent probabilities is reduced to $\sum_{r=1}^{R}(k_r - 1)$ for the initial distribution and $\sum_{r=1}^{r} k_r(k_r - 1)$ for the transition matrix. The hypothesis of (complete) independence of random effects can be directly generalized to block independence, where variables within one block are independent of variables within other blocks but otherwise arbitrarily dependent. We could also assume that some of the transition matrices ealier are equal to each other. As in Bartolucci (2006), we can also assume homogeneity (e.g. that off-diagonal elements are equal to each other), symmetry or that there are zeros in $\boldsymbol{\pi}$ or $\boldsymbol{\Pi}$. By fixing suitable structural zeros in $\boldsymbol{\Pi}$, we obtain the relevant assumption that some random effects are time-constant. A diagonal $\boldsymbol{\Pi}$ corresponds to a latent class model. Identifiability is guaranteed as soon as $\Pr(\alpha_{itrh} = \xi_{c_r rh}) > 0$ for all i, t, r and h.

Example. Suppose subjects are repeatedly measured over time and clustered in five centers (e.g. hospitals). The natural way of modelling overdispersion is to include $q_1 = 1$, subject-specific intercept and $q_2 = 4$, cluster-specific effects. The latter is associated to four dummy variables which identify the subject's cluster. It is in fact natural to assume that cluster-specific effects evolve over time *simultaneously*, and *separately* (but not necessarily independently) from the subject-specific effects. Therefore, $R = 2$ blocks shall be used. The covariates are collected in z_{it}, a binary vector with a leading one and a one in the $j + 1$-th position if the i-th subject belongs to the j-th center at time t, for $j < 5$. The subject-specific intercept has support based on k_1 values, and cluster-specific effects take one of k_2 values, with k_1 possibly different than k_2. Suppose for instance that $k_1 = 3$ and $k_2 = 2$. We then have a five dimensional random effect $\boldsymbol{\alpha}_{it}$, such that α_{it11} has support $\xi_{111}, \xi_{211}, \xi_{311}$; α_{it2h} has support $\xi_{12(h-1)}$, $\xi_{22(h-1)}$, for $h = 2, 3, 4, 5$. A total of $k1 + 4k_2 = 11$ support points are used. We model the joint distribution of random effects as $\Pr(\boldsymbol{\alpha}_{i1} = (\xi_{c_1 11}, \xi_{c_2 21}, \ldots, \xi_{c_2 24})) = \pi_{c_1, c_2}$, for $c_1 = 1, 2, 3$, $c_2 = 1, 2$. Here $\boldsymbol{\pi}$ has $k_1 k_2 = 6$ elements, corresponding to five free parameters. A 6 x6 transition matrix $\boldsymbol{\Pi}$ is analogously defined. In summary, the number of free parameters is $d + 11 + 5 + 6 * 5 = d + 46$. Under (5), the number of parameters is reduced to $d + 22$.

3. Inference

We derive inference for the general model when there are no constraints on $\boldsymbol{\xi}$, $\boldsymbol{\pi}$ or $\boldsymbol{\Pi}$. Parameters are estimated by numerical maximization of the observed log-likelihood. Maximization involves also any additional dispersion parameter ψ. Numerical maximization is

repeated from different initial starting solutions in order to increase the chances of obtaining the global optimum. Direct likelihood maximization has been argued to be more efficient than expectation-maximization in this context in a series of papers (Pinheiro and Bates, 2001, 2002; MacDonald, 2014). In models based on Markov chain assumptions, the likelihood can easily be computed through a forward recursion (Baum *et al.*, 1970; Welch, 2003), which we now extend to our modelling framework. The recursion is summarized in matrix notation, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, as

$$\boldsymbol{q}_{it}(y_{i1}, \ldots, y_{it}) = \begin{cases} \text{diag}[\boldsymbol{u}_{i1}]\boldsymbol{\pi} & \text{if } t = 1, \\ \text{diag}[\boldsymbol{u}_{it}]\boldsymbol{\Pi}'\boldsymbol{q}_{it}(y_{i1}, \ldots, y_{i,t-1}) & \text{o.w.}, \end{cases} \tag{7}$$

where $\boldsymbol{u}_{it}$ is a column vector with elements $f(y_{it}|\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_c, \boldsymbol{\beta})$. The likelihood is evaluated as $l(\boldsymbol{\theta}) = \prod_i q_{iT}(y_{i1}, \ldots, y_{iT})'\mathbf{1}$, where $\mathbf{1}$ denotes a vector of ones. The rows of $\boldsymbol{\Pi}$ and $\boldsymbol{\pi}$ are transformed through logits to account for bounded parameter spaces. When any of the proposed restrictions is adopted, the strategy is only slightly modified. An unknown permutation operator $\sigma_{rh}(\cdot)$ can be estimated through block updating, iterating maximum likelihood conditionally on the $\boldsymbol{\sigma}$ and an integer programming algorithm along the lines of Chakraborty & Chaudhury (2008). All other constraints are directly imposed by pooling, that is, the likelihood is expressed only as the set of free parameters and some of them might be repeatedly used in (7). We discuss now some additional inferential issues. First, the number of support points of each random effect can be chosen by optimizing an information criterion, like the Bayesian Information Criterion (BIC, Schwarz, 1978). Second, similarly to Bartolucci & Farcomeni (2014), the score can be computed in closed form through a first derivative of the forward recursion. This relies only on the first derivative of the log densities conditional on the fixed and random effects, which is usually available in closed form. The score is then numerically differentiated to obtain the information matrix and the standard errors. This is often more precise than numerical second derivatives of the log-likelihood at convergence. When we include unknown permutations $\sigma_{rh}(\cdot)$ by conditioning, we have that $\text{Var}(\hat{\boldsymbol{\theta}}) = \text{Var}(\text{E}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\sigma}})) + \text{E}(\text{Var}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\sigma}}))$, where the outer operators are with respect to $\hat{\boldsymbol{\sigma}}$. The first term is estimated as outlined previously, treating $\hat{\boldsymbol{\sigma}}$ as fixed. The second addend is equal to zero for all parameters except $\hat{\boldsymbol{\xi}}$, where it can be estimated with parametric bootstrap. Standard errors are then used to build Wald statistics for the regression parameters. Hypotheses about $\boldsymbol{\Pi}$ shall be tested using a likelihood ratio test (LRT) statistic, which is distributed like a chi-bar-squared in general (Bartolucci, 2006). An important issue regards dependence among random effects. In the most general formulation, groups of random effects are arbitrarily dependent and their dependence evolves over time. In order to describe this dependence, we shall use log-odds ratios and higher-order log-linear interactions. A global assessment of dependence (at a given occasion t) is obtained with the standardized total correlation (Watanabe, 1960)

$$\frac{\sum_{r=1}^{R} H(\alpha_{itr1}) - H(\boldsymbol{\alpha}_{it})}{\sum_{r=1}^{R} H(\alpha_{itr1}) - \max_{r=1}^{R} H(\alpha_{itr1})}, \tag{8}$$

where $H(\cdot)$ is Shannon's entropy. Finally, the optimal sequence of latent states (so called *global* decoding) for subject i can be estimated through the Viterbi algorithm (Viterbi, 1967).

4. Simulation study

We now simulate from a general model defined by (1), (2) and (3). The outcome is binary. We generate two predictors, one binary with probability 0.5 and one standard normal; associated with fixed effects $\boldsymbol{\beta}' = \begin{pmatrix} 1 & 1 \end{pmatrix}$. We generate $R = 3$ groups of predictors associated with random

Table 1. *Simulated data: bias and standard deviation of groups of estimates, and model deviance (**dev**)*

			$\boldsymbol{\beta}$		$\boldsymbol{\xi}_1$		$\boldsymbol{\xi}_2$		$\boldsymbol{\xi}_3$		
k_1	k_2	k_3	bias	std.dev	bias	std.dev	bias	std.dev	bias	std.dev	dev
					Proposed model						
2	2	2	0.003	0.147	0.032	0.108	0.034	0.098	0.056	0.107	2886.5
3	2	2	0.005	0.144	0.023	0.128	0.022	0.098	0.027	0.124	2972.7
2	3	2	0.006	0.143	0.024	0.133	0.033	0.141	0.042	0.123	2976.9
2	2	3	0.007	0.141	0.016	0.119	0.018	0.133	0.025	0.084	2971.0
3	3	2	0.006	0.143	0.017	0.140	0.017	0.121	0.020	0.095	3022.9
3	2	3	0.007	0.144	0.016	0.147	0.018	0.095	0.031	0.154	3014.6
2	3	3	0.006	0.149	0.022	0.146	0.022	0.141	0.033	0.155	3025.9
3	3	3	0.007	0.153	0.015	0.154	0.019	0.055	0.056	0.159	3059.5
					Restricted model						
2	2	2	0.153	0.231	0.298	0.207	0.205	0.200	0.299	0.335	2922.8
3	2	2	0.251	0.222	0.431	0.240	0.377	0.185	0.349	0.305	3022.1
2	3	2	0.102	0.239	0.221	0.194	0.197	0.221	0.284	0.273	3029.3
2	2	3	0.148	0.233	0.233	0.224	0.303	0.257	0.316	0.237	3024.2
3	3	2	0.190	0.231	0.404	0.263	0.191	0.221	0.429	0.282	3093.5
3	2	3	0.140	0.206	0.270	0.278	0.667	0.254	0.327	0.143	3087.6
2	3	3	0.084	0.178	0.128	0.138	0.333	0.199	0.363	0.127	3102.6
3	3	3	0.178	0.256	0.248	0.246	0.353	0.167	0.314	0.177	3167.8
					Classical mixed model						
2	2	2	0.393	0.275	-	-	-	-	-	-	3009.8
3	2	2	0.420	0.231	-	-	-	-	-	-	3111.9
2	3	2	0.426	0.238	-	-	-	-	-	-	3114.0
2	2	3	0.431	0.241	-	-	-	-	-	-	3111.8
3	3	2	0.433	0.203	-	-	-	-	-	-	3175.1
3	2	3	0.435	0.199	-	-	-	-	-	-	3173.1
2	3	3	0.456	0.205	-	-	-	-	-	-	3180.1
3	3	3	0.438	0.180	-	-	-	-	-	-	3208.9

effects. The first is a subject-specific intercept ($q_1 = 1$), with support points $\xi_{111} = -2$ and $\xi_{211} = 2$ when $k_1 = 2$. When $k_1 = 3$, $\xi_{311} = 0$. The second latent process is associated with a standard normal predictor ($q_2 = 1$), with the same support points as discussed earlier. The third latent process is associated with a group of $q_3 = 4$ binary predictors, which leads to cluster subjects in five approximately balanced blocks. The support points are $\boldsymbol{\xi}'_{13} = \begin{pmatrix}-2 & 2 & -2 & 2\end{pmatrix}$ and $\boldsymbol{\xi}'_{23} = \begin{pmatrix}2 & -2 & 2 & -2\end{pmatrix}$ when $k_2 = 2$. When $k_3 = 3$, $\boldsymbol{\xi}_{33}$ is constantly zero. The initial joint distribution is uniform and the joint transition matrix has $5.25/(5 + 0.25k_1k_2k_3)$ on the diagonal and equal off-diagonal elements. When $k_1 + k_2 + k_3 \geq 8$, in order to control the number of parameters, we also assume equal off-diagonal elements when estimating transition matrices.

Data is generated $B = 500$ times for $n = 500$ and $T = 5$, for each combination of $k_1 = 2, 3$, $k_2 = 2, 3$ and $k_3 = 2, 3$. Each simulated data set is used to estimate parameters of our mixed model with discrete random effects, and a restricted formulation with time-constant random effects, and a classical mixed model with Gaussian random effects. Bias and standard deviation of groups of estimates, and average model deviance (that is, minus twice the log-likelihood at convergence), are in Table 1. Deviance of our proposal is always the lowest. Mean squared error also is always smaller than the competitors, especially as a result of reduced bias. Classical mixed models are seriously biased given the random effects distribution is misspecified and T is not large.

5. Application to ADD health data

We use data from the American National Longitudinal Study of Adolescent Health (ADD Health). The study spans a period of about 15 years, recording measures of health and well-being for a sample of young american students at $T = 4$ occasions. Most subjects were 14–17

years of age in 1994, when data for the first wave were collected. The second wave was collected in 1996, then 2001 and 2008. We use the publicly available data set, which regards $n = 3087$ subjects. For each subject and at each wave, we compute the Center for Epidemiological Studies Depression Scale (CES-D), which is a gender adjusted binary indicator of depression tendencies. We investigate whether there is a health literacy effect on this outcome, that is, if health information learned at school on or before the first wave can be protective against depression tendencies. Our target predictor is the proportion, over 17 items, of health information learned at school on or before the first wave. These items include different health topics, like the foods you should and should not eat, the importance of exercise, risks of smoking, how to prevent AIDS, safety at home, how to manage stress, what if someone chokes on food, and so on. We also record race (white/non-white), an indicator of access to care (being 1 if at least once the subject was denied access to care because of lack of health insurance between the last and the present wave), an indicator of low score in the Peabody Picture Vocabulary Test taken at first wave (below the fifth percentile) and a dummy variable for each measurement occasion. Access to care is included as a proxy for socio-economic status and the Peabody score as a proxy for general cognitive abilities and efforts spent at school. There are two obvious sources of unobserved heterogeneity in this data set. The first is given by the repeated measurements on the same subjects. The second, by the fact that subjects are clustered in 131 schools, which also constitute the primary sampling units. More details can be found in Harris *et al.* (2009). In summary, $\boldsymbol{x}_{it}$ records $d = 7$ covariates associated with fixed effects, $\boldsymbol{z}_{it}$ contains a leading 1 for the subject-specific intercept and a binary 130 dimensional vector indicating the cluster of the i-th subject. Hence, $R = 2$, $q_1 = 1$ and $q_2 = 130$. Additionally, we assume the support points within the second block are k_2. Comparing all possible combinations of $k_1 = 1, \ldots, 4$ and $k_2 = 1, \ldots, 4$ through BIC, we set $k_1 = 3$ and $k_2 = 2$. In Table 2, we report the fixed effects estimates, standard errors and t-statistics for our final model, while the parameters for the random effects are in Table 3.

We conclude that after adjusting for two sources of unobserved heterogeneity and other possible confounders, the effect of health information is significant. Each additional item learned at school leads to a decrease in the probability of observing depression tendencies. This has important policy implications: schools can play a critical role in the promotion of safe and happy adolescents. The effort needed is actually minimal and involves including basic health education in the core curriculum. As often observed (Nielsen-Bohlman *et al.*, 2004), effects of health education go beyond the mere contents conveyed. Health education promotes awareness and good attitude towards taking care of oneself and others. From Table 2, we see a substantial effect of socio-economic status, with subjects being denied access to care more at risk of depression. A slight increase in risk is observed in non-whites and subjects with a low Peabody.

Table 2. *ADD Health data: fixed effects parameter estimates for a model with* $k_1 = 3$ *and* $k_2 = 2$

	$\hat{\boldsymbol{\beta}}$	s.e.	t-statistic
Access to care	1.81	0.11	17.00
Non-white	0.77	0.09	8.55
Low Peabody	1.29	0.16	8.07
Health information	–0.70	0.19	–3.66
Wave: 2 (versus wave: 1)	0.39	0.13	3.06
Wave: 3 (versus wave: 1)	1.78	0.13	13.85
Wave: 4 (versus wave: 1)	3.11	0.13	24.37

ADD Health, American National Longitudinal Study of Adolescent Health.

Table 3. *ADD Health data: random effects parameter estimates.* $\hat{\boldsymbol{\xi}}_1$, $\hat{\boldsymbol{\pi}}_1$ *and* $\hat{\boldsymbol{\Pi}}_1$ *give the support points, initial distribution and transition matrix for subject-specific parameters;* $\hat{\boldsymbol{\xi}}_2$, $\hat{\boldsymbol{\pi}}_2$ *and* $\hat{\boldsymbol{\Pi}}_2$ *the support points, initial distribution and transition matrix for cluster-specific parameters*

$\hat{\boldsymbol{\xi}}_1$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\Pi}}_1$		
–5.36	0.81	0.69	0.00	0.31
–2.42	0.01	0.01	0.99	0.00
–0.49	0.18	0.14	0.30	0.55
$\hat{\boldsymbol{\xi}}_2$	$\hat{\boldsymbol{\pi}}_2$	$\hat{\boldsymbol{\Pi}}_2$		
–5.86	0.34	0.79	0.21	
–0.61	0.66	0.52	0.48	

ADD Health, American National Longitudinal Study of Adolescent Health.

Table 4. *ADD Health data: constrained formulations for a mixed LM model with* $k_1 = 3$ *and* $k_2 = 2$. *Likelihood ratio statistics (LRT) and degrees of freedom (df) with respect to the full model, which has log-likelihood –3987.82*

Model	Description	log-likelihood	LRT	df
1	block independence	–4027.15	78.66	24
2	$\hat{\boldsymbol{\Pi}}_1$ diagonal	–4024.09	72.54	28
3	$\hat{\boldsymbol{\Pi}}_2$ diagonal	–4008.03	40.42	15
4	$\hat{\boldsymbol{\Pi}}$ diagonal	–4161.52	347.40	30

ADD Health, American National Longitudinal Study of Adolescent Health.

Finally, log-odds ratios related to wave indicators, with the first as baseline, are all positive and statistically significant: regardless of the subject's profile, depression tendencies steadily increase over time. From Table 3, we see that the three latent states for subject-specific parameters are related to low, medium and high propensity to depression. The majority of subjects are classified in the low propensity class at the first wave. Based on $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\Pi}}_1$, we see clear patterns of depression over time. Latent state 2 is an adulthood state. Young subjects are either prone or not prone to depression (states 1 and 3). In adulthood (mostly at waves 3 and 4), a third situation (state 2) appears, which is the case of subjects jumping in and out of depression (medium propensity). The fact that there are no transitions from state 1 to state 2 implies that only subjects with previous clear tendencies to depression can move to the unstable situation described by state 2. The Viterbi algorithm identifies problematic subjects, who may be contacted for intervention: 3% of the subjects start and persist in the high propensity class.

We conclude by validating our model through comparison with some constrained formulations in Table 4. LRT are computed versus the unconstrained model. The independence model (Model 1) is rejected, hence there is evidence of association between the two latent processes. The standardized total correlation is 1.4% at first wave, 30% at the second, 27% at the third and 11% at the fourth. A latent class model for cluster-specific random effects, with time-varying subject-specific intercept (Model 3) is also rejected and would be restrictive. Finally, also a

model with time-constant subject-specific intercepts (Model 2) and a full latent class model (Model 4) are rejected. Finally, we compare with more classical linear models and mixed models with two levels of Gaussian random effects, one for the subjects and one for the clusters. The residual sum of squares of our model is 526.2, while the mixed model yields 979.8 and the linear model 1268.4.

6. Conclusions

We have proposed a GLMM in which random effects can be time-varying, arbitrarily dependent and possibly high dimensional. They have a discrete support and follow a first-order Markov chain. It shall be noted that these models are often tailored for 'large n, small T' cases. Asymptotic arguments in n are rather straightforward assuming T is finite.

The matrix z_{it}, as in our example, may be high dimensional but sparse. Consequently, it can be conveniently coded in sparse form in statistical softwares, with substantial gain in computational efficiency. We have shown an application to a complex longitudinal study, investigating the relationships between health education at school and the CES-D indicator of depression. Measurement occasions in our example are not equally spaced. We had at first taken into account this fact by making the transition matrix depend on time between occasions, through a logit reparameterization of the kind $\log\left(\Pi^{(t)}_{cd}/\Pi^{(t)}_{cc}\right) = \delta_d + w_t\gamma$, for $d \neq c$; where $\mathbf{\Pi}^{(t)}$ denotes the transition matrix at the t-th time occasion, δ_d a class-specific intercept, γ a slope and w_t is the time interval between the $t-1$-th and t-th occasion. In all cases, $\hat{\gamma}$ was clearly not significant, so we used a time-homogeneous transition matrix. A substantial extension of our proposed model would be obtained if including (time-varying) random effects in a parameterization of the latent distribution, as in Altman (2007). Additionally, it could be generalized to multivariate outcomes (Bartolucci & Farcomeni, 2009), to quantile regression (Farcomeni (2012)), to informative drop-out (Bartolucci & Farcomeni (2015)) or to inhomogeneous Markov chains (Bartolucci *et al.* (2013, 2014)).

Acknowledgements

The author is grateful to an AE and two referees for constructive comments. This research uses data from ADD Health, a program of J. R. Udry, P. S. Bearman and K. M. Harris funded by grant P01-HD31921. No direct support was received from grant P01-HD31921 for this analysis.

References

Altman, R. M. (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *J. Amer. Statist. Assoc.* **102**, 201–210.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *J. Roy. Statist. Soc. Ser. B* **68**, 155–178.

Bartolucci, F. & Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *J. Amer. Statist. Assoc.* **104**, 816–831.

Bartolucci, F. & Farcomeni, A. (2014). Information matrix for hidden Markov models with covariates. *Statist. Comput.* DOI: 10.1007/s11222-014-9450-8.

Bartolucci, F. & Farcomeni, A. (2015). A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics*. DOI: 10.1111/biom.12224.

Bartolucci, F., Farcomeni, A., Pandolfi, S. & Pennoni, F. (2015). LMest: an R package for latent Markov models for categorical longitudinal data. *arXiv* **1501**, 04448.

Bartolucci, F., Farcomeni, A. & Pennoni, F. (2013). *Latent Markov models for longitudinal data*, Chapman & Hall/CRC Press, Boca Raton, FL.

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**, 433–486.
Bartolucci, F. & Lupparelli, M. (2015). Pairwise likelihood inference for nested hidden Markov chain models for multilevel longitudinal data. *J. Amer. Statist. Assoc.* DOI: 10.1080/01621459.2014.998935.
Baum, L., Petrie, T., Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**, 164–171.
Chakraborty, B. & Chaudhury, P. (2008). On an optimization problem in robust statistics. *J. Comput. Graph. Statist.* **17**, 683–702.
Diggle, P. J., Heagerty, P., Liang, K. & Zeger, S. (2002). *Analysis of longitudinal data*, OUP, Oxford.
Farcomeni, A. (2012). Quantile regression for longitudinal data based on latent Markov subject-specific parameters. *Statist. Comput.* **22**, 141–152.
Harris, K., Halpern, C., Whitsel, E., Hussey, J., Tabor, J., Entzel, P. & Udry, J. (2009). *The national longitudinal study of adolescent health: research design*. Available at: http://www.cpc.unc.edu/projects/addhealth/design [Accessed on 15 April 2014].
MacDonald, I. L. (2014). Numerical maximisation of likelihood: a neglected alternative to EM. *Int. Stat. Rev.* **82**, 296–308.
Maruotti, A. & Rydén, T. (2009). A semiparametric approach to hidden Markov models under longitudinal observations. *Statist. Comput.* **19**, 381–393.
McCulloch, C. & Searle, S. (2001). *Generalized, linear, and mixed models*, Wiley, New York.
Nielsen-Bohlman, L., Panzer, A. M. & Kindig, D. A. (eds). (2004). *Health literacy: a prescription to end confusion*, National Academies Press, Washington, DC.
Pinheiro, J. C. & Bates, D. M. (2002). *Mixed-effects models in S and S-plus*, Springer, New York.
Pinheiro, J. C., Liu, C. & Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed effects models using the multivariate T-distribution. *J. Computat. Graph. Statist.* **19**, 249–276.
Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.
Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE T. Inform. Theory* **13**, 260–269.
Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. IBM *J. Res. Dev.* **4**, 6–82.
Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Inf. Theory Soc. Newsl.* **53**, 1–13.

Received May 2014, in final form March 2015

Alessio Farcomeni, Dipartimento di Sanitá Pubblica e Malattie Infettive, Sapienza - Universitá di Roma, Piazzale Aldo Moro, 5, 00185 Rome, Italy.
E-mail: alessio.farcomeni@uniroma1.it