Biometrical Journal **xx** (2017) xx, zzz–zzz / DOI: 10.1002/bimj.200100000

Penalized estimation in latent Markov models, with application to monitoring serum Calcium levels in end-stage kidney insufficiency

Alessio Farcomeni *,[1]

[1] Department of Public Health and Infectious Diseases *(Sapienza - University of Rome)*
Piazzale Aldo Moro, 5
00185 Roma
Italy

Received zzz, revised zzz, accepted zzz

We introduce a penalized likelihood form for latent Markov models. We motivate its use for biomedical applications where the sample size is in the order of the tens, or at most hundreds, and there are only few repeated measures. The resulting estimates never break down, while spurious solutions are often obtained by maximizing the likelihood itself. We discuss model choice based on the Takeuchi Information Criterion. Simulations and a real-data application to monitoring serum Calcium levels in end-stage kidney disease are used for illustration.

Key words: penalized likelihood; bias reduction; longitudinal data; mixture models; separation.

1 Introduction

Latent Markov (LM) models (Zucchini and MacDonald, 2009; Bartolucci et al., 2013) are now a well established tool for the analysis of longitudinal data. They can be seen as finite mixtures (McLachlan and Peel, 2000) of generalized linear models (McCulloch and Searle, 2001; Farcomeni, 2015). Outcomes are possibly multidimensional (Bartolucci and Farcomeni, 2009) and extensions are available for several complex situations, including the case of informative drop-out (e.g., Bartolucci and Farcomeni (2015a)).

The difference with classical mixed effects models is that random effects have a discrete support, based on $k > 1$ support points, and that their mass probabilities are time-varying. Mass probabilities evolve over time according to a first-order (oftentimes, homogeneous) Markov chain. With longitudinal categorical outcomes, as in our example, the response can be seen as a direct or indirect measure, with error, of the latent categorical predictor. This flexible but parsimonious formulation allows the user to take into account time-varying unobserved heterogeneity. Additionally, patients can be clustered according to their latent trajectories, making random effects interpretable and directly useful for identifying subgroups. Similarly, when there are several repeated measures, one can cluster time-occasions. There are several application fields in medical statistics, often based on a single or few independent, relatively long, time series (e.g., Albert et al. (1994); Altman and Petkau (2005); Lagona et al. (2014)).

A general limitation of LM models, especially in biomedical applications, is that often a large number of observations *or* a large number of repeated measures are needed to obtain reliable estimates. In our experience LM models are very successfull in socioeconomic contexts where several hundreds subjects might be available (e.g., Farcomeni (2015)); or in ecology and engineering, where several repeated measures could be obtained albeit for one or only few subjects (e.g., Zucchini et al. (2008)). When the data matrix is neither long nor wide, the likelihood might be unbounded due to separation problems, or very

*Corresponding author: e-mail: alessio.farcomeni@uniroma1.it

unstable estimates might be obtained. In practice estimates might be unreasonably far from zero and/or huge standard errors might be observed. Additionally, there might be the occurrence of a few numerically zero or numerically one predicted probabilities for the categorical outcomes. In such cases, log-odds ratio estimates might indeed be arbitrary, as different values lead to quite similar fits. Formally, we could define break down as the occurrence of a non-positive definite information matrix at convergence, but we do not rely closely on this definition in our discussion. This problem applies also when only fixed effects are used and has been studied in depth for logistic regression models. See for instance Albert and Anderson (1984); Heinze and Schemper (2002). The problem of separation is particularly relevant in finite mixtures like LM models (e.g., García-Escudero et al. (2015) and references therein).

A typical example where the LM model could be very informative is our motivating application: serum calcium (Ca) of patients with severe kidney insufficiency was measured before surgery, and then 1 month, 1 year and 3 years after parathyroidectomy (surgical removal of parathyroids). In this example we have $n = 52$ patients and $T = 4$ occasions. Data were collected and firstly analyzed by Mazzaferro et al. (2008) with a generalized linear mixed model. The conclusions of that paper are that parathyroidectomy is not effective: the very high Ca levels drop after surgery, but then slowly increase again. A LM model would be appropriate for the data at hand. Given the long time horizon and the very few subject-specific covariates collected it is natural to expect that there might be unobserved factors influencing the outcome, and that these are changing over time. Additionally, clustering patients might be useful to identify those who benefit the most or the least from surgery. Unfortunately, the MLE of most LM model formulations is unbounded, or at least empirically breaks down, even for small numbers of latent components.

The MLE in our example breaks down regardless of the estimation strategy, that is, both for the expectation-maximization and direct maximization strategies. This happens because of the limited sample size, with only four repeated measurements.

In this paper we use a penalized likelihood approach that is particularly suitable for the cases in which the MLE might break down. The bias-corrected approach that we propose is an adaptation of the seminal work of Firth (1993), who proposed an ad-hoc inline penalty to reduce bias of maximum likelihood estimates. The resulting estimates have a lower bias than classical maximum-likelihood ones, and do not suffer from lack of convergence even with small data matrices. See also Cordeiro and McCullagh (1991) and Schaefer (1983).

With our bias-corrected approach we will be able to obtain stable estimates for our motivating example. We will be able to refine results of Mazzaferro et al. (2008) by concluding that while surgery is not effective in general, there is a small group of patients (which we are able to identify through posterior probabilities) who benefit from parathyroidectomy in the short run and might experience hypocalcemia (more easily treated in some cases) in the long run.

The penalty is based on Jeffreys prior, that is a function of the information matrix. A difficulty in LM models is that the information matrix is seldom available in closed form. We build on results of Bartolucci and Farcomeni (2015b) and demonstrate that computation of the penalized likelihood can be performed with a simple modification of the usual forward recursion for computing the observed likelihood, therefore having very little additional computational cost. The penalized likelihood can then be maximized via the usual numerical routines. In this paper we used the Nelder-Mead algorithm.

The proposed fitting strategy is particularly suited for longitudinal data arising in medical statistics. It can be seen indeed to lead to stable estimates also when the number of subjects, as in our example, is small.

The rest of the paper is as follows: in the next section we introduce our motivating data set. In Section 3 we illustrate the LM model which would be appropriate for the parathyroidectomy data. In Section 4 we outline our estimation strategy. This is demonstrated to be dramatically effective with small data matrices both in simulations in Section 5 and with the motivating data set in Section 6. Concluding remarks are given in Section 7.

2 The PTX data

NKF-K/DOQI$^{\text{TM}}$ guidelines (National Kidney Foundation, 2003) define safe serum calcium levels in the very limited range 8.5-9.5 mg/dl. While in absence of diseases the actual values can be expected to be almost always well within this interval, in kidney insufficiency calcium excretion is reduced, often leading to persistent hypercalcemia. Hypercalcemia can lead to stones, bone pain, nausea and vomiting, and even psychiatric overtones and cardiac arrhytmias. Acute hypercalcaemic crises can lead to coma. We refer here to patients with kidney insufficiency, who are also non-respondent to dyalisis. One possibility for treatment in this scenario is removal of parathyroids (parathyroidectomy, or PTX), whose hormones have the effect of increasing Ca levels. Consequently, partial or complete removal of parathyroid is expected to balance the excess Ca due to kidney insufficiency, and therefore lead to a better homeostasis of Ca serum levels.

Motivation of this work arises from an original medium-term survey on $n = 52$ subjects with severe kidney insufficiency, who are non-respondent to dyalisis, and received PTX with the aims described above. Serum calcium levels were measured before surgery, and then 1 month, 1 year and 3 years from parathyroidectomy, hence we have $T = 4$ occasions. Covariates measured include age (54.55$\pm$11.35), gender (51.9% males), dyalisis vintage (that is, length of dyalisis treatment before surgery), whose first, second and third quartile were 5, 7, and 9 years, respectively; and type of PTX (complete or partial), with 40.4% complete removal of parathyroids. Additionally we have measured basal serum parathyroid hormone (PTH) (1075.91$\pm$430.67) and Phosphate (P) levels (6.18$\pm$1.77). Both P and PTH are involved in homeostasis of Ca, but of course PTX is expected to make Ca and basal PTH levels independent.

It is natural to expect a strong unobserved heterogeneity, as many important factors were not measured and the time range of the experiment is quite long. Some of these factors (e.g., co-morbidities) might have changed during the observation period, hence unobserved heterogeneity might easily be time-varying.

Since the primary endpoint of the analysis is connected with NKF-K/DOQI$^{\text{TM}}$ ranges, we prefer to work with the response coded as an ordinal variable with three levels: below, within, and above the recommended ranges. This corresponds to hypocalcemia, good control of serum Ca, and hypercalcemia.

In Figure 1 we show the proportion of subjects with hypocalcemia, in-range values, or hypercalcemia. It can be seen that surgery is indeed effective, but only in the short term: while one month after surgery only one third of the patients still suffer from hypercalcemia, three years later this figure has raised to two thirds.

3 The model

Let Y_{it}, $i = 1, \ldots, n$ and $t = 1, \ldots, T$ denote an outcome observed for the i-th subject at time t. Let $U_{it} = 1, \ldots, k$ denote a discrete time-varying latent variable, and $\mathbf{X}_{it}$ a p-dimensional vector of possibly occasion-specific covariates.

We assume Y_{it} arises from a general exponential family. We use the generalized linear model formulation

$$g(E[y_{it}|U_{it} = u, \mathbf{X}_{it}]) = \xi_u + \mathbf{X}'_{it}\beta; \tag{1}$$

In (1) ξ_u are intercepts that are specific to the latent class, and identify a latent propensity of subjects when in latent state u.

In our application we will use global logits (Agresti and Kateri, 2011) to model a C response ordinal categorical outcome, indexed from 0 to $C - 1$. Specifically, for $y = 1, \ldots, C - 1$ we parameterize Y_{it} as

$$\log \frac{p(Y_{it} \geq y|U_{it} = u, \mathbf{X}_{it})}{p(Y_{it} < y|U_{it} = u, \mathbf{X}_{it})} = \xi_u + \alpha_y + \mathbf{X}'_{it}\beta. \tag{2}$$

It shall be noted that global logits reduce to the usual logit when $C = 2$. In (2) α_y are cut-points, where for identifiability we set $\alpha_1 = 0$. In the model above, β parameters are neither class nor category specific,

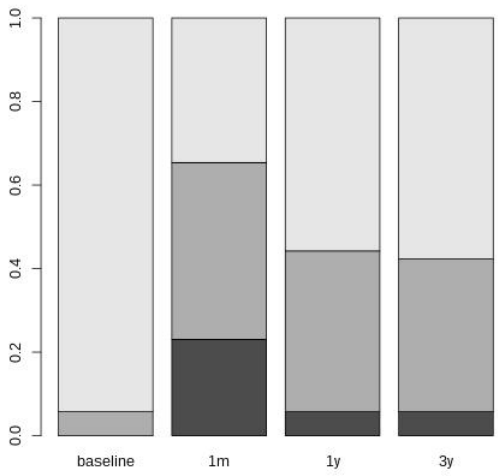

Figure 1 Proportion of subjects with hypocalcemia, in-range values, or hypercalcemia at baseline (before PTX) and 1 month, 1 year and 3 years after surgery.

and therefore summarize the global effect of the covariates. This assumption is easily relaxed if needed. As a matter of fact, the model is rather flexible and other formulations lead to minor adjustments to the inferential strategy outlined below. This is only one of the several possible formulations for LM models. In the proposed formulation all unobserved heterogeneity is summarized by latent intercepts, so that covariate effects can be more easily interpreted. To fix the ideas consider our data set where one of the predictors is serum Phosphate. We are making the reasonable assumption that unobserved confounding leads to a different propensity to hypercalcemia, but the effect of basal Phosphate is the same for everyone. The cut-points are needed when the categorical response has more than two classes. Indeed, as latent intercepts depend only on the latent state and covariate effects are homogeneous, an additional category-specific intercept (for all but the first category) is needed in order to adapt to the true underlying probability distribution.

The model is completed with assumptions for the latent variable U_{it}, where we specify an initial distribution $\Pr(U_{i1} = u) = \pi_u$, collected in a vector π, and a first-order homogeneous transition matrix Π, where $\pi_{hl} = \Pr(U_{it} = l|U_{i,t-1} = h)$ for $t > 1$.

4 Inference

4.1 Penalized Maximum Likelihood Estimation

The LM model parameters are commonly estimated through an expectation-maximization algorithm (Baum et al., 1970; Welch, 2003; Bartolucci et al., 2013), which involves forward and backward recursions at the expectation step; or by numerical maximization of the likelihood computed via a forward recursion (Turner, 2008; MacDonald, 2014; Farcomeni, 2015).

We outline here the usual forward recursion for computation of the log-likelihood. For $i = 1, \ldots, n$ let

$$a_{i1}(u) = \pi_u p(Y_{i1}|U_{i1} = u, \mathbf{X}_{i1}),$$

and for $t = 2, \ldots, T$ let

$$a_{it}(u) = p(Y_{it}|U_{it} = u, \mathbf{X}_{it}) \sum_{h=1}^{k} a_{i,t-1}(h)\pi_{hu}.$$

To avoid numerical issues when computing the sum above we proceed on the log scale, as outlined in the appendix of Farcomeni (2012). The log-likelihood can be expressed as $l(\theta) = \sum_{i=1}^{n} \log(\sum_{u=1}^{k} a_{iT}(u))$, where θ is a short-hand notation for all model parameters.

One could now maximize $l(\theta)$ numerically to obtain maximum likelihood estimates. When n and T are small, when separation (even approximately) occurs, or when variability in $\mathbf{Y}$ is small, the MLE could break down, or at least be severely biased. We propose to penalize the likelihood by means of the Jeffreys prior, that is, to obtain estimates by maximization of

$$l(\theta) + \frac{1}{2} \log |I(\theta)|, \tag{3}$$

where $I(\theta)$ denotes the observed information matrix, and $|I(\theta)|$ its determinant. The optimum of (3) can not break down, as spurious likelihood maximizers correspond to an infinite $\log |I(\theta)|$. Several approaches have appeared in the literature for computing the information matrix of LM models, e.g., Hughes (1997), Lystig and Hughes (2002), Bartolucci and Farcomeni (2015b). All these approaches are convenient as they only invove additional recursions which can run in parallel with the one above. In this paper we compute the score analytically through the first derivative of the forward recursion above (Lystig and Hughes, 2002). The information matrix is then obtained numerically through a numerical first derivative of the score. This approach is general and convenient, and has been argued to provide accurate standard errors estimates (e.g., Bartolucci and Farcomeni (2009)). After running the appropriate recursions for a given value of θ one can simply compute (3) after obtaining the determinant of $I(\theta)$ numerically.

Maximization of (3) can be performed numerically as well. For convenience we proceed after transformation of the constrained parameters. Specifically we use a local logit parameterization of π and Π as $\log(\pi_u/\pi_1)$ for $u = 2, \ldots, k$ and $\log(\pi_{hl}/\pi_{hh})$ for $h = 1, \ldots, k$ and $l \neq h$.

In all cases, in order to increase the odds of finding the global optimum, we compared the results obtained from a few different starting solutions, where the first one is obtained deterministically by initializing regression parameters at zero, uniform initial distribution and symmetric transition matrix as in the simulation study (see below); and the other ones are obtained at random.

4.2 Model choice and standard errors

Note that if parameter estimates correspond to the maximum penalized likelihood estimator (pMLE), the model is essentially misspecified and commonly used information criteria, like AIC and BIC, are not formally correct for model choice.

It is worth recalling that AIC is defined as $-2l(\hat{\theta}) + 2g$, where g is the number of free parameters. The penalty parameter $2g$ corresponds to an elegant estimate of the bias that one has when using the same data for estimating both the deviance (i.e., $-2l(\hat{\theta})$) and the parameter θ. If the model is misspecified (for instance, θ is estimated based on an objective function that is different than the likelihood) then the deviance is still a good summary for model evaluation, but bias does not correspond to $2g$ anymore. See for instance Konishi and Kitagawa (1996), who explicitly treat the case of penalized likelihoods.

It is now well known that under misspecification the penalty of the Akaike Information Criterion should be substituted with

$$2\mathrm{tr}(s(\theta)s(\theta)'I(\theta)^{-1})\big|_{\theta=\hat{\theta}},$$

where $s(\theta)$ is the score, $\hat{\theta}$ the parameter estimate, and tr is the trace operator. This gives Takeuchi Information Criterion (TIC), derived in Takeuchi (1976), which is a generalization of AIC for misspecified models.

In our setting the score vector and information matrix are available, after recursion, in closed form and hence computing TIC is straightforward. The score is computed with a (parallel) recursion alike that of the log-likelihood and information matrix (Hughes, 1997; Lystig and Hughes, 2002; Bartolucci and Farcomeni, 2015b).

An issue with TIC is that when the sample size is small, which is the relevant situation under study, it might be unstable. This is easily detected through a sensitivity analysis (that is, by varying score or information matrix entries by a small amount and checking that TIC is not too different). In the context of LM models, TIC is used to choose the number of latent states k. This makes the use of the following strategy particularly convenient: fix a tolerance ε (e.g., $\varepsilon = 1e-3$). Call $TIC(k)$ the TIC associated with a model with k latent states. For each $k = 2, \ldots, (k_{\max} - 1)$, add and remove ε from each entry of the score and information matrix, separately, and check whether the order between $TIC(k)$ and $TIC(k-1)$ or between $TIC(k)$ and $TIC(k+1)$ is reversed. If so, $TIC(k)$ is not stable. For $k = 1$ check only if the order between $TIC(k)$ and $TIC(k+1)$ is reversed. Similarly, for $k = k_{\max}$ check only whether the order between $TIC(k)$ and $TIC(k-1)$ is reversed.

In case TIC is unstable we suggest to complement/confirm model choice via cross-validation. This is done by randomly splitting the sample in a training (e.g., 80% of the observations) and a test set (the remaining ones). The model is estimated on the training set and the error evaluated by checking prediction on the test set. The procedure is repeated several times and the final average prediction error is used for model choice.

Standard errors for the pMLE can not be evaluated as usual via the inverse of the observed information matrix. As outlined above, the model is misspecified. Consequently, a sandwich estimator must be computed as

$$I(\hat{\theta})'(s(\hat{\theta})s(\hat{\theta})')^{-1}I(\hat{\theta}).$$

All quantities above are already computed for computation of $\hat{\theta}$, hence there is minimal additional computational effort.

5 Simulation study

In order to illustrate the benefits of maximizing (3) instead of the observed likelihood for estimation of model parameters we set up a simulation study.

We do so based on model (2). We let $n = \{50, 100\}$, $T = \{3, 4\}$, $C = 3$, $k = \{2, 3\}$, $p = 2$. When $k = 2$, latent intercepts are $\xi_1 = -8$ and $\xi_2 = -2$ and α_y are equally spaced between 0 and 10, regardless of C. When $k = 3$, $\xi_1 = -8$, $\xi_2 = -3$ and $\xi_3 = 2$ while α_y are equally spaced between 0 and 6. The two covariates are independently sampled from a Bernoulli with probability 50% and a standard Gaussian distribution, respectively. We fix $\beta = (1\ 1)$, a uniform initial distribution π, and $\pi_{lh} \propto 0.2$ if $h \neq l$ and 1.2 otherwise. The transition matrix is then scaled so that the rows sum to 1.

For each scenario we generate $B = 1000$ data sets, and estimate parameters based on usual maximization of the likelihood and of the penalized likelihood (3). A non-positive definite observed information matrix at convergence is identified as the occurrence of failure (due to lack of local identifiability, for instance).

In Table 1 we report the failing probability for maximum likelihood estimation and, for non-failing iterates, the average root mean squared error (RMSE), additionally averaged over groups of parameters.

It can be seen that in almost all scenarios the probability of failure for the MLE is very large. Additionally, the RMSE of the pMLE is always substantially smaller (due to reduced bias) than that of the MLE, with the exception of the RMSE for π when $n = 100$. We speculate this happens as the EM for the MLE is initialized from the true value for π, and therefore it might be a sign that the EM is not actually moving much from the initial solution for this parameter.

The advantage of using the pMLE might vanish with larger n or T, but for small data matrices it is substantial. It could be noted that the breakdown rate depends on the way latent intercepts and cutoffs were fixed at data generation.

In order to validate TIC we also have run an additional simulation study in which k was not known. To do so, for each scenario we have generated the data, estimated all possible models for $k = 1, 2, 3, 4$ and chosen the best k through TIC, AIC, and BIC. We report in Table 2 the proportion of times among

Table 1 Simulation results. $|I(\hat{\theta}_{MLE})| \leq 0$ is the probability of non-positive definite information matrix at convergence for the MLE. RMSE: root mean squared error, pMLE: penalized MLE. Results are based on $B = 1000$ replicates

				RMSE α		RMSE ξ		RMSE β		RMSE π		RMSE Π	
n	T	k	$\lvert I(\hat{\theta}_{MLE})\rvert \leq 0$	MLE	pMLE	MLE	pMLE	MLE	pMLE	MLE	pMLE	MLE	pMLE
50	3	2	0.78	5.46	1.88	10.45	5.51	1.63	0.49	0.23	0.12	0.21	0.06
50	3	3	0.78	20.61	1.77	35.12	6.71	10.84	0.81	0.11	0.13	0.13	0.09
50	4	2	0.63	4.84	1.91	9.68	4.89	0.77	0.48	0.23	0.10	0.18	0.05
50	4	3	0.52	15.67	1.74	27.78	6.69	8.93	0.77	0.09	0.11	0.11	0.08
100	3	2	0.42	4.33	2.30	9.10	4.76	0.43	0.34	0.20	0.29	0.18	0.08
100	3	3	0.32	13.59	1.70	24.43	6.68	6.27	0.66	0.08	0.09	0.10	0.08
100	4	2	0.20	3.91	2.28	8.67	4.73	0.54	0.34	0.18	0.30	0.14	0.08
100	4	3	0.06	7.78	1.68	16.08	6.70	3.31	0.65	0.07	0.08	0.08	0.08

$B = 1000$ replicates that each value of k was selected by each information criterion. In bold we report the proportion for the number of latent states used for data generation.

Table 2 Simulation results. Proportion over $B = 1000$ replicates that $k = 1, 2, 3, 4$ was chosen using TIC, AIC or BIC.

			TIC				AIC				BIC			
n	T	k	1	2	3	4	1	2	3	4	1	2	3	4
50	3	2	0.0	**90.0**	7.8	2.2	0.0	**51.3**	48.7	0.0	0.0	**94.0**	6.0	0.0
50	3	3	0.0	71.9	**14.3**	13.8	0.0	49.5	**40.9**	9.6	0.0	72.2	**26.9**	0.9
50	4	2	0.0	**79.5**	16.2	4.3	0.0	**28.8**	71.1	0.1	0.0	**78.3**	21.7	0.0
50	4	3	0.0	34.2	**38.2**	27.6	0.0	51.6	**35.3**	13.1	0.0	76.4	**21.9**	1.7
100	3	2	0.0	**95.3**	3.3	1.4	0.0	**93.9**	6.1	0.0	0.0	**99.5**	0.5	0.0
100	3	3	0.0	25.6	**48.1**	26.3	0.0	34.9	**36.8**	28.3	0.0	90.1	**9.7**	0.2
100	4	2	0.0	**87.7**	8.6	3.7	0.0	**80.4**	19.6	0.0	0.0	**98.5**	1.5	0.0
100	4	3	0.0	23.6	**42.1**	34.3	0.0	24.5	**37.2**	38.3	0.0	60.5	**33.5**	6.0

Results indicate that no method works well with small data matrices. When $n = 50$ and $T = 3$ AIC might be better than the other competitors, possibly due to a slight instability of TIC penalty. When $n = 100$ or $T = 4$ TIC seems to outperform the other information criteria as it chooses $k = 2$ with high probability when $k = 2$ and when $k = 3$ the true k is selected with a sensibly larger probability than the other methods. Suppose now that, as in our example, $n = 50$, $T = 4$, and $k = 3$ is selected with TIC. If we assume that either $k = 2$ or $k = 3$ with equal prior probabilities, the (posterior) probability that the data generating k is $k = 3$ can be computed as $0.382/(0.382 + 0.162) = 70.2\%$

6 Data Analysis

For $k = 1, \ldots, 5$ we have fit the model proposed in Section 4 to the PTX data. We have done so both by maximization of the likelihood, and by maximization of the penalized likelihood. As covariates we use time dummies and baseline Phosphate serum levels. We did not obtain significance of any other covariate, even after checking several possible specifications of the other model parameters. This is reasonable as we do not really expect dependence on gender, age, dyalisis vintage, and type of surgery (complete vs partial).

In Table 3 we report log-likelihood at convergence for the MLE and pMLE for $k = 1, \ldots, 5$. The best model according to the TIC criterion is based on $k = 3$. Note that for fixed k the MLE achieves a likelihood

that is slightly higher than the pMLE in all cases. This is natural as pMLE is based on a penalized objective function.

Table 3 PTX data analysis. Log-likelihood at convergence and TIC for MLE and penalized likelihood estimator (pMLE).

	MLE		pMLE	
k	lik	AIC	lik	TIC
1	-156.7	325.4	-156.7	313.7
2	-144.6	307.2	-143.7	297.7
3	-137.4	302.8	-140.4	290.6
4	-135.5	309.1	-143.1	307.1
5	-129.7	319.3	-143.4	305.0

We note that TIC for the pMLE is stable with these data, based on a sensitivity analysis as outlined in the previous section. We report that $TIC(5)$ could be changed by a relatively large amount by adding or removing small values to the score entries, but in no case it could be lower than $TIC(3)$. Hence the conclusions based on TIC are not affected by absolute lack of stability of $TIC(5)$, as there is relative stability.

In Table 4 we compare parameter estimates for $k = 1, \ldots, 5$, obtained with the MLE and pMLE. It can be seen that MLE estimates break down for $k \geq 3$, while pMLE ones do not. In Table 4, $\beta_1, \beta_2, \beta_3, \beta_4$ regard coefficients for second vs first measurement occasion, third vs first, fourth vs first, and Phosphate serum levels before surgery, respectively. As expected there is a positive association between serum Calcium and serum Phosphate levels. The parameter estimates reported in Table 4 in correspondence of $k = 3$ for the pMLE are all significant at the 5% level after Wald test.

Table 4 PTX data analysis. Parameter estimates with MLE and pMLE for $k = 1, \ldots, 5$.

	MLE					pMLE				
Effect	k					k				
	1	2	3	4	5	1	2	3	4	5
α_2	2.21	4.27	26.24	13.82	217.02	2.16	3.93	4.71	5.23	4.43
β_1	-3.60	-6.28	-59.24	-37.97	-359.78	-3.42	-5.90	-6.02	-8.99	-7.90
β_2	-2.54	-4.24	-35.82	-30.62	-213.30	-2.39	-4.18	-3.57	-5.84	-5.73
β_3	-2.47	-4.30	-43.92	-32.59	-276.88	-2.32	-4.32	-3.38	-5.65	-5.10
β_4	0.08	0.33	0.37	1.32	13.04	0.08	0.35	0.21	-0.30	0.15
ξ_1	2.81	7.37	33.73	36.24	-13.36	2.65	1.79	-3.20	-3.11	-2.55
ξ_2		2.17	-1.23	-1.28	97.43		7.21	6.79	2.37	4.13
ξ_3			60.33	58.60	174.22			7.21	9.21	9.47
ξ_4				25.80	340.19				10.04	10.59
ξ_5					419.44					16.36

For the chosen model with $k = 3$ we additionally report the initial probability vector, which corresponds to $(0.42\ 0.03\ 0.55)$, and the hidden transition matrix

$$\begin{pmatrix} 0.86 & 0.02 & 0.12 \\ 0.08 & 0.31 & 0.62 \\ 0.11 & 0.07 & 0.81 \end{pmatrix}.$$

Given that there are $C = 3$ levels for the outcome, interpretation of the parameter estimates is not straightforward. We proceed by comparison of $k = 3$ plots, where the predicted probabilities of each category at each time occasion, conditionally on the latent state, are reported. This is given in Figure 2,

where basal Phosphate levels are set at the median. The light rectangle indicates the expected proportion with hypercalcemia, the dark rectangle the expected proportion with hypocalcemia, and the middle one the expected proportion of patients with in-range serum Ca levels. This should be compared with Figure 1, which is the marginal distribution of the outcome stratified by measurement occasion, and therefore a weighted average of estimates in Figure 2.

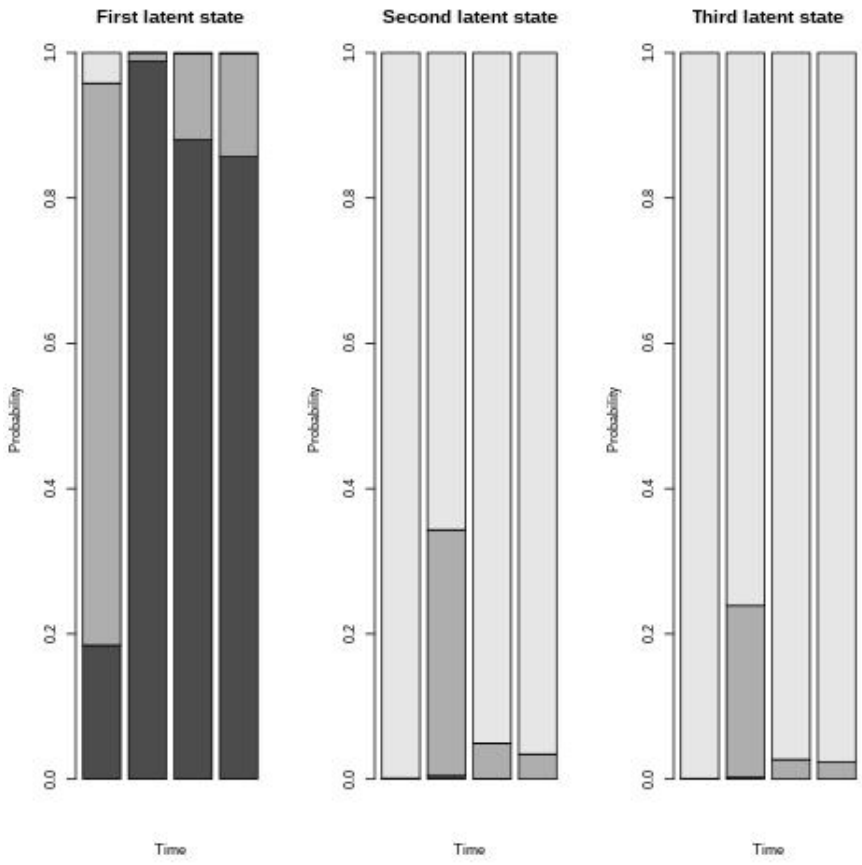

Figure 2 PTX data. Predicted probabilities of each category at each time occasion, conditionally on each of $k = 3$ latent states, for basal P levels at the observed median. The light rectangle indicates the expected proportion with hypercalcemia, the dark rectangle the expected proportion with hypocalcemia, and the middle one the expected proportion of patients with in-range serum Ca levels. Estimates are based on pMLE.

It can be seen that a first latent state is associated with relatively good baseline Ca levels, and subsequent hypocalcemia. The second latent state is associated with the best 1 month profile, and mostly hypercalcemia later; the third latent state is slightly worse than the second one. The first latent state is arguably very well separated from the other two. Notably, also the second and third are significantly separated as, for instance, the LRT for comparing $k = 2$ with $k = 3$ is rejected. These results clearly confirm the findings of Mazzaferro et al. (2008): surgery might not be effective in the long run, as at 3 and 5 years patients with in-range serum Ca levels are a minority. Additionally, recall that with our LM models subjects are allowed to move from one latent state to another. Regardless of their latent state at baseline, we can identify patients in first latent state at 1 month, 3 and 5 years as patients who actually have been lead to

hypocalcemia by surgery. After adjusting for baseline Phosphate levels and regardless of the initial latent state, we obtain that 36% of patients undergoing PTX will have this latent trajectory. On the other hand, the expected proportion of patients in the second latent state at 1 month, regardless of previous states, is 12%. Hence we can conclude that while PTX is generally ineffective, a small fraction of patients *will* benefit from this surgical procedure in the short run and an even higher fraction will experience hypocalcemia in the long run, instead of hypercalcemia. Depending on co-morbidities and the general health of the patient, nephrologists might more easily treat hypocalcemia (first latent state) rather than hypercalcemia (second and third latent state). All these patients are worth of further investigation. While we could not explain with the available covariates why their surgery was more effective, a closer examination of the differences of these patients with respect to the remaining cohort might generate hypotheses for the identification of determinants of outcomes of PTX.

An additional issue is that measurement occasions are not equally spaced in time, while a homogeneous hidden transition matrix is assumed by our models. It shall be said that transitions are expected to occur quickly after surgery, and then more and more slowly. The time points where chosen by nephrologists approximately targeting homogeneity, as a matter of fact. In order to formally verify homogeneity one should include time between occasions as a covariate after a logit parameterization of the hidden transition matrix, which would be beyond the scope of this work. In order to provide some evidence in favor of homogeneity of the hidden transition matrix we have computed the residuals of the pMLE when $k = 3$, and estimated a pMLE for the residuals. The model with $k = 1$ for the residuals has TIC 138.2, for $k = 2$ it has 154.0 and for $k = 3$ 163.8. Hence a model with $k = 1$ is chosen with TIC, indicating that even with a homogeneous hidden transition matrix there is no residual unobserved heterogeneity.

We conclude by noting that basically no classical mixed-effect model formulation was useful, with this data, for similarly identifying the subgroup of patients for whom the treatment might be effective.

7 Conclusions

We have outlined a penalized likelihood approach for LM models. We have focused on a specific model formulation for illustrating purposes, but generalization to any LM model is straightforward. We have showed that the pMLE is particularly useful when the data matrix is small and/or there might be separation. Likelihood of separation is increased with the number of latent states k, as one subgroup of subjects might be separable. The pMLE is computationally convenient and, most importantly, it does not break down. It is directly connected to Bayesian Maximum-A-Posteriori estimates based on Jeffreys' prior. Several other penalties are possible, of course. In this work we have worked with the Jeffreys prior in order to mimick Firth's penalized likelihood approach. In fact, Firth's approach reduces to penalizing through the Jeffreys' prior in full exponential families.

Implementation of our approach is straightforward through already available software, e.g., through function `est_lm_cov_manifest` in `R` package `LMest` (Bartolucci et al., 2015). We have referenced different strategies for computing the information matrix exactly. In general cases (e.g., multivariate outcomes, mixed effects latent Markov models) one can numerically differentiate the score, as in Bartolucci and Farcomeni (2009).

We have additionally discussed model choice, which can be based on TIC and sensitivity analysis. If sensitivity analysis fails, cross-validation is recommended despite being potentially time-consuming.

Finally, we have discussed a challenging data set in nephrology, where we have seen that PTX in kidney insufficiency might in general not be as effective as would be desired. Additionally, we have identified a subgroup of patients whose PTX can be labeled as successful in the short run, and another group who might experience hypocalcemia in the long run. This has been done based on the posterior probabilities of the latent states, which is what makes LM models worth using in this application. What makes these patients different from the other ones deserves further investigation. This must be based on additional

biomarkers (e.g., serum vitamin D levels, eGFR, etc.) which are not available at the moment. These additional biomarkers could be used to parameterize the manifest or latent distributions, therefore contributing to observed heterogeneity and/or latent transitions.

Supplementary Materials

R code is available as Supporting Information on the journal's web page, together with source code to reproduce all the simulation results.

Acknowledgments

The author thanks the Department of Clinical Science, Sapienza - University of Rome, for permission to use the PTX data, and an AE and two referees for kind and constructive comments.

Conflict of Interest *The author declares no conflict of interest*

References

Agresti, A. and Kateri, M. (2011). *Categorical data analysis*. Springer.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1–10.

Albert, P.; McFarland, H.; Smith, M. and Frank, J. (1994). Time series for modelling counts from a relapsing-remitting disease: Application to modelling disease activity in multiple sclerosis. *Statistics in Medicine* **13**, 453–466.

Altman, R. M. and Petkau, A. J. (2005). Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine* **24**, 2335–2344.

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. and Farcomeni, A. (2015a). A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics* **71**, 80–89.

Bartolucci, F. and Farcomeni, A. (2015b). Information matrix for hidden Markov models with covariates. *Statistics and Computing* **25**, 515–526.

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

Bartolucci, F.; Farcomeni, A. and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC Press, Boca Raton, FL.

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.

Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society (Series B)* **53**, 629–643.

Farcomeni, A. (2012). Quantile Regression for longitudinal data based on latent Markov subject-specific parameters. *Statistics and Computing* **22**, 141–152.

Farcomeni, A. (2015). Generalized linear mixed models based on latent Markov heterogeneity structures. *Scandinavian Journal of Statistics* **42**, 1127–1135.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

García-Escudero, L. A.; Gordaliza, A.; Matrán, C. and Mayo-Iscar, A. (2015). Avoiding Spurious Local Maximizers in Mixture Modeling. *Statistics and Computing* **25**, 619–633.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Hughes, J. (1997). Computing the observed information in the hidden Markov model using the EM algorithm. *Statistics & Probability Letters* **32**, 107–114.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

Lagona, F.; Jdanov, D. and Shkolnikova, M. (2014). Latent time-varying factor in longitudinal analysis: a linear mixed hidden Markov model for heart rates. *Statistics in Medicine* **33**, 4116–4134.

Lystig, T. C. and Hughes, J. (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics* **11**, 678–689.

MacDonald, I. L. (2014). Numerical Maximisation of Likelihood: A Neglected Alternative to EM. *International Statistical Review* **82**, 296–308.

Mazzaferro, S.; Pasquali, M.; Farcomeni, A.; Vestri, A.; Filippini, A.; Romani, A. M.; Barresi, G. and Pugliese, F. (2008). Parathyroidectomy as a Therapeutic Tool for Targeting the Recommended NKF-K/DOQI Ranges for Serum Calcium, Phosphate and Parathyroid Hormone in Dialysis Patients. *Nephrology Dialysis Transplantation* **23**, 2319–2323.

McCulloch, C. and Searle, S. (2001). *Generalized, linear, and mixed models*. Wiley, New York.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

National Kidney Foundation (2003). K/DOQI clinical practice guidelines: bone metabolism and disease in chronic kidney disease. *American Journal of Kidney Diseases* **42**, S1–S201.

Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine* **2**, 71–78.

Takeuchi, K. (1976). Distribution of information Statistics and Criteria for Adequacy of Models. *Mathematical Science* **153**, 12–18.

Turner, R. (2008). Direct maximization of the likelihood of a hidden Markov model. *Computational Statistics and Data Analysis* **52**, 4147–4160.

Welch, L. R. (2003). Hidden Markov Models and the Baum-Welch Algorithm. *IEEE Information Theory Society Newsletter* **53**, 1–13.

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

Zucchini, W.; Raubenheimer, D. and MacDonald, I. L. (2008). Modeling time series of animal behaviour by means of a latent-state model with feedback. *Biometrics* **64**, 807–815.

