Stat Comput
DOI 10.1007/s11222-014-9450-8

Information matrix for hidden Markov models with covariates

Francesco Bartolucci · Alessio Farcomeni

Received: 1 May 2013 / Accepted: 18 January 2014

Abstract For a general class of hidden Markov models that may include time-varying covariates, we illustrate how to compute the observed information matrix, which may be used to obtain standard errors for the parameter estimates and check model identifiability. The proposed method is based on the Oakes' identity and, as such, it allows for the exact computation of the information matrix on the basis of the output of the expectation-maximization (EM) algorithm for maximum likelihood estimation. In addition to this output, the method requires the first derivative of the posterior probabilities computed by the forward-backward recursions introduced by Baum and Welch. Alternative methods for computing exactly the observed information matrix require, instead, to differentiate twice the forward recursion used to compute the model likelihood, with a greater additional effort with respect to the EM algorithm. The proposed method is illustrated by a series of simulations and an application based on a longitudinal dataset in Health Economics.

Keywords EM algorithm · Forward-backward recursions · Oakes' identity · Standard errors

F. Bartolucci (✉)
Department of Economics, University of Perugia,
Via A. Pascoli, 20, 06123 Perugia, Italy
e-mail: bart@stat.unipg.it

A. Farcomeni
Department of Public Health and Infectious Diseases,
Sapienza - University of Rome,
Piazzale Aldo Moro, 5,
00185 Roma, Italy
e-mail: alessio.farcomeni@uniroma1.it

1 Introduction

Hidden Markov (HM) models have been initially developed in the literature on stochastic processes as extensions for measurement errors of the standard Markov chain model; for one of the oldest contributions about these models see Baum and Petrie (1966). Then, these models have received considerable attention in the time-series literature, due to their wide applicability and ease of interpretation (for an up-to-date review see Zucchini and MacDonald 2009). HM models have also found an increasing popularity for the analysis of multiple series of data (e.g., Turner et al. 1998), and in the related context of longitudinal data (Bartolucci et al. 2013b), in which short series of repeated observations are available for many sample units.

Though algorithms for the direct maximization of the model likelihood have attracted recent interest in the HM literature, see Turner (2008) and references therein, the main tool for maximum likelihood (ML) estimation of the parameters of an HM model is the expectation-maximization (EM) algorithm, which is based on certain forward-backward recursions. These are commonly known as Baum–Welch recursions (Baum et al. 1970; Welch 2003). The EM algorithm may be implemented with a reasonable effort and is very stable to reach convergence, whereas direct maximization algorithms are in general less stable for HM models. However, the EM algorithm does not provide, as a by-result, the standard errors for the parameter estimates. This is because it uses neither the observed nor the expected information matrix. The information matrix is also important to check local identifiability of the model through its rank; see McHugh (1956) and Goodman (1974) among others.

In the statistical literature, different techniques have been proposed to compute the information matrix, at least approximately, on the basis of the output of the EM algorithm; for

Published online: 13 February 2014

a review see McLachlan and Krishnan (2008). In particular, the Louis (1982)'s method is based on the missing information principle as defined by Orchard and Woodbury (1972). According to this principle, the observed information matrix can be expressed as the difference between two matrices corresponding to the *complete information* and the *missing information* due to the unobserved variables. Another decomposition was proposed by Oakes (1999) and is based on an explicit formula for the second derivative matrix of the model log-likelihood. This formula involves the first derivative of the conditional expectation of the score of the complete data log-likelihood, given the observed data.

In the literature on HM models, the decomposition proposed by Louis (1982) has been applied by different authors to compute the information matrix; see, in particular, Hughes (1997) and Turner et al. (1998). See also Bartolucci and Farcomeni (2009) for a related method based on the numerical derivative of the score. However, at least to our knowledge, the decomposition of Oakes (1999) has not yet been applied to obtain the observed information matrix in the present context.

In this article, we show how to use the Oakes' identity for a general class of HM models that include time-varying covariates. The proposed method uses the incomplete data information matrix, which is produced by the EM algorithm, and a correction matrix computed on the basis of the first derivative of the posterior probabilities obtained from the Baum–Welch recursions. There are two clear advantages: on one hand, the method is exact; on the other hand, little computational effort is needed beyond the usual EM computations.

We have to clarify that methods for exactly obtaining the information matrix already exist in the HM literature (Khan 2003; Lystig and Hughes 2002; Cappé et al. 2005; Turner 2008). They are based on the first and second derivatives of the forward recursion to obtain the model likelihood (Baum et al. 1970), or on similar approaches. However, these methods do not employ the output of the EM algorithm and are a more natural choice in connection with direct maximization of the model likelihood. Furthermore, computing the second derivatives of each element of the recursion in the presence of covariates, especially when they are continuous, may be a daunting task.

The method proposed in this article does not require to differentiate twice the forward recursion, with a certain computational advantage. It is also rather obvious that, in order to obtain standard errors for the parameter estimates, we can alternatively use a parametric bootstrap method, as described in Zucchini and MacDonald (2009), or a nonparametric bootstrap method (Efron and Tibshirani 1993). However, bootstrapping may be computationally heavy and, in any case, does not allow us to check for local identifiability directly.

The remainder of the paper is organized as follows. In the next section we review some preliminary notions. The proposed method to compute the observed information matrix of the model on the basis of Oakes' identity is illustrated in Sect. 3. Finally, in Sect. 4 we illustrate the approach through a series of simulations and an application based on a longitudinal dataset about Health Economics.

All the methods proposed in this article have been implemented in a series of `R` functions, which rely on a `Fortran` code to speed up the execution and are available upon request.

2 Preliminaries about hidden Markov models

In the following, we briefly review the class of HM models with possible time-varying covariates that are of interest for the proposed developments. We also describe the Baum–Welch recursions to compute the manifest distribution of the data and the posterior distribution of the latent variables and we outline maximum likelihood estimation by the EM algorithm. Note that we initially discuss in more detail the case of a single series of data and we show how this formulation can be used to deal with multiple series and longitudinal data. The adopted notation is rather similar to that used in Bartolucci et al. (2013b), to which we refer the reader for details.

2.1 Model formulation

With a single series of data, let T denote the number of occasions of observation and suppose that, for $t = 1, \ldots, T$, we observe a vector of r response variables, denoted by $\boldsymbol{Y}^{(t)} = (Y_1^{(t)}, \ldots, Y_r^{(t)})$, and vectors of covariates $\boldsymbol{W}^{(t)}$ and $\boldsymbol{X}^{(t)}$ entering in the so-called *measurement component* and *latent component* of the model, respectively. All these variables are collected in the vectors $\tilde{\boldsymbol{W}}$, $\tilde{\boldsymbol{X}}$, and $\tilde{\boldsymbol{Y}}$. The HM model with covariates is formulated by assuming a latent process $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(T)})$ that follows a first-order Markov chain with state space $\{1, \ldots, k\}$ and initial and transition probabilities possibly depending on the covariates in $\tilde{\boldsymbol{X}}$. Under the assumption of *local independence*, the observed random vectors $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}$ are conditionally independent given the underlying latent process and the covariates in $\tilde{\boldsymbol{W}}$. This assumption leads to a strong simplification of the model, and it can be relaxed by assuming that each $\boldsymbol{Y}^{(t)}$ also depends on $\boldsymbol{Y}^{(t-1)}$. These assumptions are illustrated by the path diagram in Fig. 1.

Let $f_A(\cdot)$ denote the probability (or density) for the distribution of a random variable A, and $f_{B|A}(\cdot|\cdot)$ its conditional counterpart for two random variables A and B; this notation extends to random vectors in a natural way. With reference to the HM framework introduced above, we let

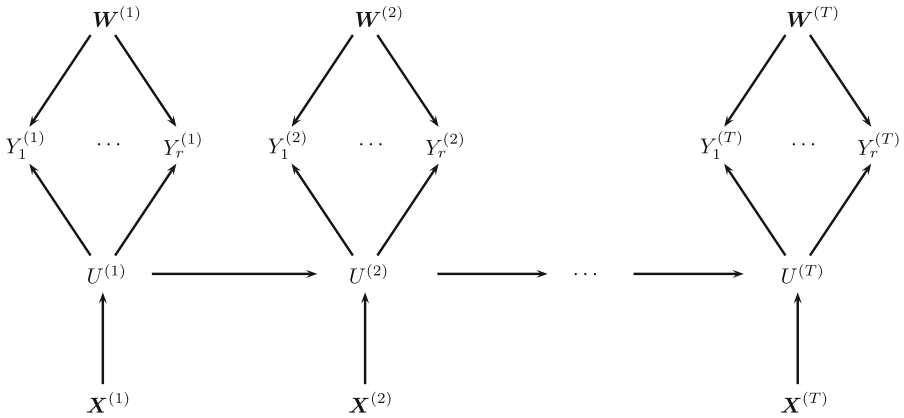

Fig. 1 Path diagram for the HM model with covariates

$$\phi^{(t)}(\boldsymbol{y}|u,\boldsymbol{w}) = f_{\boldsymbol{Y}^{(t)}|U^{(t)},\boldsymbol{W}^{(t)}}(\boldsymbol{y}|u,\boldsymbol{w}), \quad t = 1,\ldots,T, \; u = 1,\ldots,k,$$
$$\pi(u|\boldsymbol{x}) = f_{U^{(1)}|\boldsymbol{X}^{(1)}}(u|\boldsymbol{x}), \quad u = 1,\ldots,k,$$
$$\pi^{(t)}(u|\bar{u},\boldsymbol{x}) = f_{U^{(t)}|U^{(t-1)},\boldsymbol{X}^{(t)}}(u|\bar{u},\boldsymbol{x}), \quad t = 2,\ldots,T, \; \bar{u},u = 1,\ldots,k.$$

In the above expressions, by u and $\bar{u}$ we denote possible realizations of the latent variables, by $\boldsymbol{w}$ we denote a realization of $\boldsymbol{W}^{(t)}$, by $\boldsymbol{x}$ a realization of $\boldsymbol{X}^{(t)}$, and by $\boldsymbol{y} = (y_1,\ldots,y_r)$ a realization of $\boldsymbol{Y}^{(t)}$. In practice, formulating assumptions on the measurement component amounts to express $\phi^{(t)}(\boldsymbol{y}|u,\boldsymbol{w})$ as a function of the covariates depending on suitable parameters to be estimated. Accordingly, formulating assumptions on the latent component amounts to suitably parametrize the initial probabilities $\pi(u|\boldsymbol{x})$ and the transition probabilities $\pi^{(t)}(u|\bar{u},\boldsymbol{x})$.

Concerning the measurement component, the model is specified by a parametrization of $\phi^{(t)}(\boldsymbol{y}|u,\boldsymbol{w})$ of canonical GLM type (McCullagh and Nelder 1989), so that the model may be used with continuous, binary, or count response variables. In particular, in the univariate case in which there is a single response variable $Y^{(t)}$ for each time occasion ($r = 1$), with $\eta^{(t)}(u,\boldsymbol{w})$ being a transformation of the mean $\mu^{(t)}(u,\boldsymbol{w}) = E(Y^{(t)}|U^{(t)} = u, \boldsymbol{W}^{(t)} = \boldsymbol{w})$ based on a suitable link function, we assume that

$$\eta^{(t)}(u,\boldsymbol{w}) = (\boldsymbol{a}_{u\boldsymbol{w}}^{(t)})'\boldsymbol{\alpha}. \tag{1}$$

In the previous expression, $\boldsymbol{a}_{u\boldsymbol{w}}^{(t)}$ is a column vector depending on the value of $U^{(t)}$ and that of $\boldsymbol{W}^{(t)}$ and $\boldsymbol{\alpha}$ is the corresponding vector of regression parameters. Just to clarify, if we assume that each variable $Y^{(t)}$ has a Poisson conditional distribution, then the link function is based on the logarithm and we may assume that

$$\eta^{(t)}(u,\boldsymbol{w}) = \log\mu^{(t)}(u,\boldsymbol{w}) = \alpha_{1u} + \boldsymbol{w}'\boldsymbol{\alpha}_2,$$

that may be reformulated as in (1).

To cover the case of categorical variables with more than two categories, we also consider a multinomial logit parameterization (Agresti 2002). In particular denoting by c the number of response categories, labelled from 0 to $c-1$, we consider the vector $\boldsymbol{\eta}^{(t)}(u,\boldsymbol{w})$, with elements equal to the logits

$$\eta^{(t)}(y|u,\boldsymbol{w}) = \log\frac{\phi^{(t)}(y|u,\boldsymbol{w})}{\phi^{(t)}(0|u,\boldsymbol{w})}, \quad y = 1,\ldots,c-1,$$

and then we assume that

$$\boldsymbol{\eta}^{(t)}(u,\boldsymbol{w}) = \boldsymbol{A}_{u\boldsymbol{w}}^{(t)}\boldsymbol{\alpha}, \tag{2}$$

where $\boldsymbol{A}_{u\boldsymbol{w}}^{(t)}$ is a design matrix depending on u and $\boldsymbol{w}$. This link function may be simply inverted as follows

$$\boldsymbol{\phi}^{(t)}(u,\boldsymbol{w}) = \frac{\exp[\boldsymbol{G}_{1c}\boldsymbol{\eta}^{(t)}(u,\boldsymbol{w})]}{\mathbf{1}_c'\exp[\boldsymbol{G}_{1c}\boldsymbol{\eta}^{(t)}(u,\boldsymbol{w})]},$$

where $\boldsymbol{\phi}^{(t)}(u,\boldsymbol{w})$ is a column vector with elements $\phi^{(t)}(y|u,\boldsymbol{w})$, $y = 0,\ldots,c-1$, $\boldsymbol{G}_{hc}$ is obtained by removing the h-th column from the matrix $\boldsymbol{I}_c$, the identity matrix of dimension c, and $\mathbf{1}_c$ is a column vector of c ones. It is also important to note that, with categorical variables (also binary), the model may be formulated using as parameters the conditional response probabilities $\phi(y|u) = \phi^{(t)}(y|u,\boldsymbol{w})$, $t = 1,\ldots,T, u = 1,\ldots,k, y = 0,\ldots,c-1$, which are collected in the vector $\boldsymbol{\phi}$ in a suitable order. We consider this formulation for the application discussed in Sect. 4.2. In this way, the covariates are ruled out from the measurement model, but it is still convenient to adopt a parametrization of the type above, which is based on logits and in which $\boldsymbol{\alpha}$ substitutes $\boldsymbol{\phi}$, in order to avoid certain constraints on the parameter space. In this case, the above design vectors and matrices have elements only equal to 0 or 1.

In the multivariate case ($r > 1$), it is common to assume that the response variables $Y_1^{(t)},\ldots,Y_r^{(t)}$ collected in $\boldsymbol{Y}^{(t)}$ are conditionally independent given $U^{(t)}$ and $\boldsymbol{W}^{(t)}$, so that

$$\phi^{(t)}(\boldsymbol{y}|u,\boldsymbol{w}) = \prod_{j=1}^{r}\phi_j^{(t)}(y_j|u,\boldsymbol{w}), \tag{3}$$

where $\phi_j^{(t)}(y_j|u, \boldsymbol{w})$ is referred to the conditional distribution of $Y_j^{(t)}$ given $U^{(t)}$ and $\boldsymbol{W}^{(t)}$. Then, the same parametrization as above may be used for each of these conditional distributions. These parametrizations rely on a GLM formulation of type (1), based on vectors $\boldsymbol{a}_{ju\boldsymbol{w}}^{(t)}$, or on a formulation of type (2), based on design matrices $\boldsymbol{A}_{ju\boldsymbol{w}}^{(t)}$, for the case of categorical response variables. In both cases, a common vector $\boldsymbol{\alpha}$ of regression parameters is used. Moreover, in the case of categorical response variables, the vector of the conditional probabilities of $Y_j^{(t)}$ given $U^{(t)}$ and $\boldsymbol{W}^{(t)}$ is denoted by $\boldsymbol{\phi}_j^{(t)}(u, \boldsymbol{w})$.

In the multivariate case, it may be also reasonable to assume that the response variables in $\boldsymbol{Y}^{(t)}$ are not conditionally independent given $U^{(t)}$ and $\boldsymbol{W}^{(t)}$, allowing then for a form of *contemporary dependence*. For this aim, we may adopt a single link function for the entire multivariate distribution $\phi^{(t)}(\boldsymbol{y}|u, \boldsymbol{w})$ depending on a single vector of parameters $\boldsymbol{\eta}^{(t)}(u, \boldsymbol{w})$. Such an approach has been proposed for categorical response variables by Bartolucci and Farcomeni (2009). However, this approach may be difficult to adopt when outcomes are of mixed nature, and therefore it is in general assumed that (3) holds. Finally, it is interesting to note that, by including in each vector of covariates $\boldsymbol{W}^{(t)}$ the lagged response variables, that is the vector $\boldsymbol{Y}^{(t-1)}$, we can easily relax the hypothesis of local independence allowing for *serial dependence*; see Bartolucci and Farcomeni (2009) for details.

Regarding the latent component, we can parameterize the initial and transition probabilities of the latent Markov chain depending on the covariates in $\tilde{\boldsymbol{X}}$. For the initial probabilities we assume that

$$\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{B}_{\boldsymbol{x}}$ is an appropriate design matrix depending on the covariates in $\boldsymbol{X}^{(1)}$, $\boldsymbol{\beta}$ is a vector of parameters, and $\boldsymbol{\lambda}(\boldsymbol{x})$ is the vector of multinomial logits having the first as reference categories, that is,

$$\lambda(u|\boldsymbol{x}) = \log \frac{\pi(u|\boldsymbol{x})}{\pi(1|\boldsymbol{x})}, \quad u = 2, \ldots, k.$$

The typical assumption formulated in this way is

$$\lambda(u|\boldsymbol{x}) = \beta_{1u} + \boldsymbol{x}'\boldsymbol{\beta}_{2u}, \tag{5}$$

which may be reformulated as in (4). For the transition probabilities, we assume that

$$\boldsymbol{\rho}^{(t)}(\bar{u}, \boldsymbol{x}) = \boldsymbol{C}_{\bar{u}\boldsymbol{x}}^{(t)}\boldsymbol{\gamma}, \tag{6}$$

where $\boldsymbol{\rho}^{(t)}(\bar{u}, \boldsymbol{x})$ contains the multinomial logits

$$\rho^{(t)}(u|\bar{u}, \boldsymbol{x}) = \log \frac{\pi^{(t)}(u|\bar{u}, \boldsymbol{x})}{\pi^{(t)}(\bar{u}|\bar{u}, \boldsymbol{x})}, \quad u = 1, \ldots, k, \ u \neq \bar{u},$$

$\boldsymbol{C}_{\bar{u}\boldsymbol{x}}^{(t)}$ depends on the covariates in $\boldsymbol{X}^{(t)}$ and $\boldsymbol{\gamma}$ is a vector of logistic regression parameters. A natural parametrization that may be assumed in this case is

$$\rho^{(t)}(u|\bar{u}, \boldsymbol{x}) = \gamma_{1\bar{u}u} + \boldsymbol{x}'\boldsymbol{\gamma}_{2\bar{u}u}. \tag{7}$$

However, the model complexity may be strongly reduced by assuming that the regression parameters do not depend on the current latent state, so that $\boldsymbol{\gamma}_{2\bar{u}u} = \boldsymbol{\gamma}_{2u}, \bar{u} = 1, \ldots, k$. Both parametrizations may be formulated as in (6) with the parameters $\gamma_{1\bar{u}u}$ and $\boldsymbol{\gamma}_{2\bar{u}u}$ (or $\boldsymbol{\gamma}_{2u}$) included in $\boldsymbol{\gamma}$ and the corresponding design matrix structured accordingly.

As mentioned at the beginning of this section, in certain contexts we observe n "parallel" sequences of data which are assumed to be independent. This case is of main interest for the models here illustrated and is typical of longitudinal studies in which n is much larger than T. On the other hand, the case of multiple time series governed by a common latent process may be casted in the theory illustrated above, with $Y_j^{(t)}$ referred to the observation at occasion t for time series j.

In order to clarify the above arguments, suppose that we repeatedly observe n sample units at T time occasions. Then, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, we denote the vector of the r response variables by $\boldsymbol{Y}_i^{(t)} = (Y_{i1}^{(t)}, \ldots, Y_{ir}^{(t)})$ and the vectors of covariates by $\boldsymbol{W}_i^{(t)}$ and $\boldsymbol{X}_i^{(t)}$. All these variables are collected in the individual vectors $\tilde{\boldsymbol{W}}_i, \tilde{\boldsymbol{X}}_i$, and $\tilde{\boldsymbol{Y}}_i, i = 1, \ldots, n$. In this context, the HM model with covariates, which is also called *latent Markov model* (Bartolucci et al. 2013b), is formulated by assuming n individual latent processes. The latent process assumed for sample unit $i = 1, \ldots, n$ is denoted by $\boldsymbol{U}_i = (U_i^{(1)}, \ldots, U_i^{(T)})$ and, on the basis of this process, the same assumptions as for a single series of data are formulated. In particular, the assumption of local independence is formulated by requiring that the vectors $\boldsymbol{Y}_i^{(1)}, \ldots, \boldsymbol{Y}_i^{(T)}$ are conditionally independent given $\boldsymbol{U}_i$ and $\tilde{\boldsymbol{W}}_i$. Obviously, the assumption that every latent process $\boldsymbol{U}_i$ follows a Markov chain distribution of first order is also replicated. Finally, in the case of longitudinal data, the measurement component and the latent component are expressed for each sample unit i in the same way as illustrated above. We only need to specify design matrices in (2), (4), and (6) which are individual specific, but nothing changes in terms of model formulation and the same vectors of parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are common to all units.

2.2 Manifest and posterior distributions

Regardless of the specific model formulation, the assumption of local independence implies that for the conditional distribution of $\tilde{\boldsymbol{Y}}$, given $\boldsymbol{U}$ and $\tilde{\boldsymbol{W}}$, we have

$$f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U},\tilde{\boldsymbol{W}}}(\tilde{\boldsymbol{y}}|\boldsymbol{u}, \tilde{\boldsymbol{w}}) = \prod_{t=1}^{T} \phi^{(t)}(\boldsymbol{y}^{(t)}|u^{(t)}, \boldsymbol{w}^{(t)}),$$

for any realization $\tilde{\boldsymbol{y}}$ of $\tilde{\boldsymbol{Y}}$ (with subvectors $\boldsymbol{y}^{(t)}$), $\tilde{\boldsymbol{w}}$ of $\tilde{\boldsymbol{W}}$ (with subvectors $\boldsymbol{w}^{(t)}$), and $\boldsymbol{u}$ of $\boldsymbol{U}$ (with elements $u^{(t)}$). Moreover, since we assume that the latent process follows a first-order Markov chain, we have that

$$f_{\boldsymbol{U}|\tilde{\boldsymbol{X}}}(\boldsymbol{u}|\tilde{\boldsymbol{x}}) = \pi(u^{(1)}|\boldsymbol{x}^{(1)}) \prod_{t=2}^{T} \pi^{(t)}(u^{(t)}|u^{(t-1)}, \boldsymbol{x}^{(t)}),$$

for any realization $\tilde{\boldsymbol{x}}$ of $\tilde{\boldsymbol{X}}$ (with subvectors $\boldsymbol{x}^{(t)}$).

Now let $f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}) = f_{\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{Z}}}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})$ denote the probability mass (or density) function for the manifest distribution of $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{Z}}$, where $\tilde{\boldsymbol{Z}}$ denotes the vector of the covariates obtained by the union of the vectors $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{X}}$, so that redundant elements are avoided, and $\tilde{\boldsymbol{z}}$ is a corresponding realization (with subvectors $\tilde{\boldsymbol{z}}^{(t)}$ made of the union of $\boldsymbol{w}^{(t)}$ and $\boldsymbol{x}^{(t)}$). We have that

$$f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}) = \sum_{\boldsymbol{u}} f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U},\tilde{\boldsymbol{W}}}(\tilde{\boldsymbol{y}}|\boldsymbol{u}, \tilde{\boldsymbol{w}}) f_{\boldsymbol{U}|\tilde{\boldsymbol{X}}}(\boldsymbol{u}|\tilde{\boldsymbol{x}}),$$

where the sum $\sum_{\boldsymbol{u}}$ is extended to all possible k^T configurations $\boldsymbol{u}$ of the latent process. The posterior distribution of the latent process, which corresponds to the conditional distribution of $\boldsymbol{U}$ given $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{Y}}$, is denoted by $q(\boldsymbol{u}|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}}) = f_{\boldsymbol{U}|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{Z}}}(\boldsymbol{u}|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}})$ and has expression

$$q(\boldsymbol{u}|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}}) = \frac{f_{\tilde{\boldsymbol{Y}}|\boldsymbol{U},\tilde{\boldsymbol{W}}}(\tilde{\boldsymbol{y}}|\boldsymbol{u}, \tilde{\boldsymbol{w}}) f_{\boldsymbol{U}|\tilde{\boldsymbol{X}}}(\boldsymbol{u}|\tilde{\boldsymbol{x}})}{f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})}.$$

In order to compute efficiently the above densities and probabilities, we employ the Baum–Welch recursions. Let

$$l^{(t)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}) = f_{U^{(t)},\boldsymbol{Y}^{(1)},\ldots,\boldsymbol{Y}^{(t)}|\boldsymbol{Z}^{(1)},\ldots,\boldsymbol{Z}^{(t)}} (u, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(t)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(t)}), \quad t = 1, \ldots, T.$$

The forward recursion is initialized with $l^{(1)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}) = \pi(u|\boldsymbol{x}^{(1)})\phi^{(1)}(\boldsymbol{y}^{(1)}|u, \boldsymbol{w}^{(1)})$, and it is based on the following step

$$l^{(t)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}) = \sum_{\bar{u}=1}^{k} l^{(t)}(\bar{u}, u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}), \quad u = 1, \ldots, k, \tag{8}$$

to be performed for $t = 2, \ldots, T$, where

$$l^{(t)}(\bar{u}, u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}) = l^{(t-1)}(\bar{u}, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})\pi^{(t)}(u|\bar{u}, \boldsymbol{x}^{(t)})\phi^{(t)}(\boldsymbol{y}^{(t)}|u, \boldsymbol{w}^{(t)}).$$

At the end of the forward recursion, the manifest distribution is simply obtained as

$$f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}) = \sum_{u=1}^{k} l^{(T)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}).$$

The other recursion introduced by Baum and Welch is a backward recursion, which allows us to obtain the posterior distribution of every latent state and of every pair of consecutive latent states. Let

$$m^{(t)}(\tilde{\boldsymbol{y}}|\bar{u}, \tilde{\boldsymbol{z}}) = f_{\boldsymbol{Y}^{(t+1)},\ldots,\boldsymbol{Y}^{(T)}|U^{(t)},\boldsymbol{Z}^{(t+1)},\ldots,\boldsymbol{Z}^{(T)}} (\boldsymbol{y}^{(t+1)}, \ldots, \boldsymbol{y}^{(T)}|\bar{u}, \boldsymbol{z}^{(t+1)}, \ldots, \boldsymbol{z}^{(T)}),$$

for $t = 1, \ldots, T-1$ and $\bar{u} = 1, \ldots, k$. This recursion is initialized with $m^{(T)}(\tilde{\boldsymbol{y}}|\bar{u}, \tilde{\boldsymbol{z}}) = 1$ and it is based on the following steps:

$$m^{(t)}(\tilde{\boldsymbol{y}}|\bar{u}, \tilde{\boldsymbol{z}}) = \sum_{u=1}^{k} m^{(t)}(u, \tilde{\boldsymbol{y}}|\bar{u}, \tilde{\boldsymbol{z}}), \quad u = 1, \ldots, k,$$

to be performed in reverse order (i.e., from $t = T-1$ to $t = 1$), where

$$m^{(t)}(u, \tilde{\boldsymbol{y}}|\bar{u}, \tilde{\boldsymbol{z}}) = m^{(t+1)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{z}})\pi^{(t+1)}(u|\bar{u}, \boldsymbol{x}^{(t+1)}) \phi^{(t+1)}(\boldsymbol{y}^{(t+1)}|u, \boldsymbol{w}^{(t+1)}).$$

From the results above, and also using the results from the forward recursion, we obtain the following posterior probabilities:

$$q^{(t)}(u|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}}) = f_{U^{(t)}|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{Z}}}(u|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}}) = \frac{l^{(t)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})m^{(t)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{z}})}{f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})}, \quad u = 1, \ldots, k, \tag{9}$$

for $t = 1, \ldots, T$, and

$$q^{(t)}(\bar{u}, u|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}}) = f_{U^{(t-1)},U^{(t)}|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{Z}}}(\bar{u}, u|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}}) = \frac{l^{(t)}(\bar{u}, u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})m^{(t)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{z}})}{f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})}, \quad \bar{u}, u = 1, \ldots, k, \tag{10}$$

for $t = 2, \ldots, T$.

The developments proposed in the present article require to compute the first derivative of the logarithm of the above posterior probabilities with respect to the overall parameter vector $\boldsymbol{\theta}$, which collects $\boldsymbol{\alpha}$ (involved in the measurement component) and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (involved in the latent component). In fact, following Oakes (1999), these derivatives are required for the adjustment of the information matrix corresponding to the expected value of the complete data log-likelihood, coming from the EM algorithm, so as to obtain the observed information matrix. We simply have that

$$\frac{\partial \log q^{(t)}(u|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}})}{\partial \boldsymbol{\theta}} = \frac{\partial \log l^{(t)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})}{\partial \boldsymbol{\theta}} + \frac{\partial \log m^{(t)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{z}})}{\partial \boldsymbol{\theta}} - \frac{\partial \log f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})}{\partial \boldsymbol{\theta}},$$

$$\frac{\partial \log q^{(t)}(\bar{u}, u|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}})}{\partial \boldsymbol{\theta}} = \frac{\partial \log l^{(t)}(\bar{u}, u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})}{\partial \boldsymbol{\theta}} + \frac{\partial \log m^{(t)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{z}})}{\partial \boldsymbol{\theta}} - \frac{\partial \log f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})}{\partial \boldsymbol{\theta}},$$

where the derivatives of $\log l^{(t)}(u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})$, $\log l^{(t)}(\bar{u}, u, \tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})$, $\log m^{(t)}(\tilde{\boldsymbol{y}}|u, \tilde{\boldsymbol{z}})$, and $\log f(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}})$ may be computed as clarified in Appendix 1.

Finally, it is useful to recall that, in order to avoid numerical instability that may occur with large values of T, it may be necessary to implement suitable normalizations; see Scott (2002). This also motivated Lystig and Hughes (2002) to use a recursion for computing the model likelihood that is similar to the Baum–Welch forward recursion, but it does not suffer from numerical instability. See also Khreich et al. (2010) and Bartolucci and Pandolfi (2014) for an up to date review of alternatives to the Baum–Welch recursions, and Farcomeni (2012) for a different strategy based on logarithm summation.

2.3 Maximum likelihood estimation

For the sake of generality we consider the case of n independent series of multivariate data, so that we suppose to observe the vectors of responses $\tilde{\boldsymbol{y}}_i$, with elements $y_{ij}^{(t)}$, and the corresponding vectors of covariates $\tilde{\boldsymbol{z}}_i$ with subvectors $\boldsymbol{z}_i^{(t)}$ given by the union of $\boldsymbol{w}_i^{(t)}$ and $\boldsymbol{x}_i^{(t)}$. The model log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \log f_i(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{z}}_i),$$

where $f_i(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{z}}_i)$ is the manifest distribution of $\tilde{\boldsymbol{y}}_i$, which is computed by the forward recursion illustrated above.

In order to maximize $\ell(\boldsymbol{\theta})$, the EM algorithm employs the *complete data log-likelihood*, which has expression:

$$\begin{aligned}
\ell^*(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \ell_i^*(\boldsymbol{\theta}), \\
\ell_i^*(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{u=1}^{k} d_{iu}^{(t)} \sum_{j=1}^{r} \log \phi_j^{(t)}(y_{ij}^{(t)}|u, \boldsymbol{w}_i^{(t)}) \\
&\quad + \sum_{u=1}^{k} d_{iu}^{(1)} \log \pi(u|\boldsymbol{x}_i^{(1)}) \\
&\quad + \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k} d_{i\bar{u}u}^{(t)} \log \pi^{(t)}(u|\bar{u}, \boldsymbol{x}_i^{(t)}), \qquad (11)
\end{aligned}$$

where $d_{iu}^{(t)}$ and $d_{i\bar{u}u}^{(t)}$ are unit-specific indicator variables. In particular, $d_{iu}^{(t)}$ is equal to 1 if unit i is in latent state u at occasion t and to 0 otherwise, whereas $d_{i\bar{u}u}^{(t)}$ is equal to 1 if this unit moves from latent state $\bar{u}$ to state u at occasion t and to 0 otherwise.

Starting from a point in the parameter space, the EM algorithm maximizes the model log-likelihood by alternating two steps until convergence. At the E-step, we compute the conditional expected value of each indicator variable involved in the complete log-likelihood given the observed data $\tilde{\boldsymbol{y}}_i$, the covariates $\tilde{\boldsymbol{z}}_i$, and the current value of the parameter vector, denoted by $\bar{\boldsymbol{\theta}}$. This simply amounts to compute the posterior probabilities of every latent variable $U_i^{(t)}$ and every pair $(U_i^{(t-1)}, U_i^{(t)})$. These posterior probabilities are computed as in (9) and (10) and are denoted by $q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})$ and $q^{(t)}(\bar{u}, u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})$, respectively, where the argument $\bar{\boldsymbol{\theta}}$ recalls that they depend on the parameter vector obtained at the previous iteration of the algorithm.

By substituting these quantities in (11) we obtain the conditional expected value of the complete data log-likelihood, which may be decomposed as

$$Q(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}) = Q_1(\boldsymbol{\alpha}|\bar{\boldsymbol{\theta}}) + Q_2(\boldsymbol{\beta}|\bar{\boldsymbol{\theta}}) + Q_3(\boldsymbol{\gamma}|\bar{\boldsymbol{\theta}}).$$

At the M-step, $\boldsymbol{\theta}$ is updated by separately maximizing each component of $Q(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}})$. A Newton-Raphson algorithm is used for this aim; hence we need the first and second derivatives of each component. About the first component $Q_1(\boldsymbol{\alpha}|\bar{\boldsymbol{\theta}})$, these derivatives may be expressed as follows:

$$\begin{aligned}
\frac{\partial Q_1(\boldsymbol{\alpha}|\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{u=1}^{k} q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}}) \\
&\quad \times \sum_{j=1}^{r} \frac{\partial \log \phi_j^{(t)}(y_{ij}^{(t)}|u, \boldsymbol{w}_i^{(t)})}{\partial \boldsymbol{\alpha}}, \\
\frac{\partial^2 Q_1(\boldsymbol{\alpha}|\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'} &= \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{u=1}^{k} q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}}) \\
&\quad \times \sum_{j=1}^{r} \frac{\partial^2 \log \phi_j^{(t)}(y_{ij}^{(t)}|u, \boldsymbol{w}_i^{(t)})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'},
\end{aligned}$$

where the derivatives of $\log \phi_j^{(t)}(y_{ij}^{(t)}|u, \boldsymbol{w}_i^{(t)})$ are defined in Appendix 2. In a similar way we can express the first and second derivatives of $Q_2(\boldsymbol{\beta}|\bar{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\beta}$ and of $Q_3(\boldsymbol{\gamma}|\bar{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\gamma}$. In this case we need the derivatives of $\log \pi(u|\boldsymbol{x}^{(1)})$ and $\log \pi(u|\bar{u}, \boldsymbol{x}^{(t)})$ which are again given in Appendix 2.

As typically happens for latent variable and mixture models, the likelihood function may be multimodal. In particular, the EM algorithm could converge to a mode of the likelihood which does not correspond to the global maximum. In order to increase the chance of reaching the global maximum, the EM must be initialized properly. In particular, we suggest to use a deterministic starting solution and to compare the value of the log-likelihood at convergence with values obtained starting from randomly chosen initial values. For a related multi-start strategy for mixture models see Berchtold (2004). We refer the reader to Bartolucci et al. (2013b) for additional details.

3 Observed information matrix

In this section, we illustrate how to implement Oakes' identity to obtain the observed information matrix under the modeling assumptions formulated in Sect. 2.1. This can be done with a reduced effort using the output of the EM algorithm and the recursions described above.

First of all, recall that the score vector and the observed information matrix for the log-likelihood $\ell(\boldsymbol{\theta})$ are defined as

$$\boldsymbol{s}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \text{and} \quad \boldsymbol{J}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'},$$

respectively. Moreover, with reference to a general latent variable model that is estimated on the basis of an EM algorithm of the type illustrated in Sect. 2.3, we have

$$\boldsymbol{s}(\boldsymbol{\theta}) = \left. \frac{\partial Q(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right|_{\bar{\boldsymbol{\theta}}=\boldsymbol{\theta}}.$$

Based on this result, Oakes (1999) derived the following identity that involves two components:

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\left\{ \left. \frac{\partial^2 Q(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\bar{\boldsymbol{\theta}}=\boldsymbol{\theta}} + \left. \frac{\partial^2 Q(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}'} \right|_{\bar{\boldsymbol{\theta}}=\boldsymbol{\theta}} \right\}. \tag{12}$$

The first component is the second derivative of the conditional expected value of the complete data log-likelihood given the observed data. This component is directly provided by the EM algorithm. The second component involved in (12) is the first derivative of the score, for the same expected log-likelihood, with respect to the current value of the parameters.

It is straightforward to see that the first component in (12) is a block-diagonal matrix defined as follows:

$$\frac{\partial^2 Q(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \operatorname{diag}\left(\frac{\partial^2 Q_1(\boldsymbol{\alpha}|\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}}, \frac{\partial^2 Q_2(\boldsymbol{\beta}|\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}}, \frac{\partial^2 Q_3(\boldsymbol{\gamma}|\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}} \right).$$

In order to compute the second component in (12) we need the first derivatives of the expected values in (9) and (10) with respect to $\bar{\boldsymbol{\theta}}$, so as to obtain

$$\frac{\partial^2 Q(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\theta}'} = \left(\frac{\partial^2 Q_1(\boldsymbol{\alpha}|\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\alpha}'}, \frac{\partial^2 Q_2(\boldsymbol{\beta}|\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\beta}'}, \frac{\partial^2 Q_3(\boldsymbol{\gamma}|\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\gamma}'} \right).$$

In this regard we have to consider that

$$\frac{\partial q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}}} = q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}}) \frac{\partial \log q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}}}$$

and the derivative of $q^{(t)}(\bar{u}, u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})$ with respect to $\bar{\boldsymbol{\theta}}$ may be defined in a similar way.

The derivatives of $\log q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})$ and $\log q^{(t)}(\bar{u}, u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})$ may be computed as clarified in Sect. 2.2. We therefore have that

$$\frac{\partial^2 Q_1(\boldsymbol{\alpha}|\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\alpha}'} = \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{u=1}^{k} q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}}) \times \frac{\partial \log q^{(t)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}}} \sum_{j=1}^{r} \frac{\partial \log \phi_j^{(t)}(y_{ij}^{(t)}|u, \boldsymbol{w}_i^{(t)})}{\partial \boldsymbol{\alpha}'},$$

$$\frac{\partial^2 Q_2(\boldsymbol{\beta}|\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \sum_{u=1}^{k} q^{(1)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}}) \times \frac{\partial \log q^{(1)}(u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}}} \frac{\partial \log \pi(u|\boldsymbol{x}_i^{(1)})}{\partial \boldsymbol{\beta}'},$$

$$\frac{\partial^2 Q_3(\boldsymbol{\gamma}|\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}} \partial \boldsymbol{\gamma}'} = \sum_{i=1}^{n} \sum_{t=2}^{T} \sum_{\bar{u}=1}^{k} \sum_{u=1}^{k} q^{(t)}(\bar{u}, u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}}) \times \frac{\partial \log q^{(t)}(\bar{u}, u|\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{z}}_i; \bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}}} \frac{\partial \log \pi^{(t)}(u|\bar{u}, \boldsymbol{x}_i^{(t)})}{\partial \boldsymbol{\gamma}'}.$$

The observed information at the ML estimate $\hat{\boldsymbol{\theta}}$ can be used to obtain the standard errors and check model identifiability. In particular, the standard errors are obtained by computing the square root of the elements in the main diagonal of $\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}$. Local identifiability is checked through the rank of $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$; that this matrix is of full rank is of course required in order to compute its inverse.

Note that the standard errors obtained as above are referred to the ML estimate of the parameter vector $\boldsymbol{\theta}$. However, we can simply express the standard errors for the estimate of a transformation of these parameters by the delta method. For instance, consider the case in which, as in the application described in Sect. 4.2, the covariates do not directly affect the conditional distribution of the response variables and these are categorical. Then it may be interesting to compute the standard errors for the conditional probabilities $\phi(y|u)$ collected in the vector $\boldsymbol{\phi}$. In order to obtain standard errors for those quantities, we simply need to compute the derivatives of $\boldsymbol{\theta}$, with respect to $\boldsymbol{\phi}, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$, as we show in Appendix 2. Once the derivatives above have been computed, they are used to build a block diagonal matrix (the last two blocks correspond to an identity matrix). According to the delta method, this matrix pre and post multiplies $\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}$. The square root of the elements in the main diagonal of the resulting matrix contains the standard errors for $\hat{\phi}(y|u)$.

4 Empirical illustration

In this section we illustrate the proposed approach by a brief simulation study and an application in the context of Health Economics, which is based on the Health and Retirement Study (HRS) dataset described below.

Table 1 Simulation results in terms of length and coverage of Wald confidence intervals (at 95 % level) computed on the basis of the standard errors obtained by the proposed method under scenarios in which covariates do not affect (no cov.) or affect (cov.) the initial and/or the transition probabilities

Initial prob.	Transition prob.	T	n	$\boldsymbol{\phi}$		$\boldsymbol{\beta}$		$\boldsymbol{\gamma}$	
				Length	Coverage	Length	Coverage	Length	Coverage
No cov.	No cov.	5	2,000	0.4100	0.9468	0.0434	0.9449	1.0124	0.9595
No cov.	No cov.	5	4,000	0.2879	0.9487	0.0305	0.9509	0.6493	0.9547
No cov.	No cov.	10	2,000	0.3305	0.9530	0.0259	0.9500	0.4701	0.9467
No cov.	No cov.	10	4,000	0.2331	0.9519	0.0183	0.9463	0.3291	0.9423
Cov.	No cov.	5	2,000	0.5025	0.9520	0.0350	0.9474	0.7492	0.9500
Cov.	No cov.	5	4,000	0.3524	0.9547	0.0247	0.9471	0.5170	0.9513
Cov.	No cov.	10	2,000	0.4563	0.9533	0.0239	0.9474	0.4343	0.9478
Cov.	No cov.	10	4,000	0.3211	0.9544	0.0169	0.9538	0.3039	0.9528
No cov.	Cov.	5	2,000	0.4252	0.9477	0.0447	0.9437	1.0636	0.9413
No cov.	Cov.	5	4,000	0.2992	0.9513	0.0314	0.9478	0.7248	0.9548
No cov.	Cov.	10	2,000	0.3466	0.9447	0.0273	0.9438	0.5379	0.9548
No cov.	Cov.	10	4,000	0.2443	0.9545	0.0193	0.9458	0.3772	0.9448
Cov.	Cov.	5	2,000	0.5248	0.9527	0.0363	0.9468	0.8221	0.9523
Cov.	Cov.	5	4,000	0.3674	0.9533	0.0256	0.9427	0.5700	0.9450
Cov.	Cov.	10	2,000	0.4767	0.9517	0.0252	0.9518	0.4891	0.9542
Cov.	Cov.	10	4,000	0.3355	0.9462	0.0178	0.9572	0.3432	0.9498

4.1 Simulation study

We considered a simulation setting that recalls that of the application based on the HRS data. In particular, we assumed a single response variable with $c = 5$ categories, labeled from zero to four, that is observed at $T = 5, 10$ occasions for a sample of size $n = 2000, 4000$. Moreover, for every sample unit and time occasion we considered two covariates that are independently generated from a standard normal distribution.

As in the application, we assumed an HM model with covariates not affecting the conditional response probabilities, $k = 2$ latent states, and conditional response probabilities chosen as follows:

$$\phi(y|1) = (5 - y)/15, \quad \phi(y|2) = (y + 1)/15, \quad y = 0, \ldots, 4,$$

so that the first latent state corresponds to a higher tendency to respond by one of the first categories, and the second state corresponds to a higher tendency to respond by one of the last categories. We also considered different data generating mechanisms, in which the covariates may or not affect the initial probabilities of the latent process and its transition probabilities according to the logit parametrizations in (5) and (7). In particular, when the covariates do not affect the initial probabilities we let $\boldsymbol{\beta}_{22} = (0, 0)'$ and when the covariates affect these probabilities we let $\boldsymbol{\beta}_{22} = (1, -1)'$; in both cases we fixed $\beta_{12} = 0$. In a similar way, when the covariates do not affect the transition probabilities, we let $\boldsymbol{\gamma}_{212} = \boldsymbol{\gamma}_{221} = (0, 0)'$ and when the covariates affect these probabilities we let $\boldsymbol{\gamma}_{212} = \boldsymbol{\gamma}_{221} = (1, -1)'$; in both cases we fixed $\gamma_{112} = \gamma_{121} = -\log 9$, so as to include a high level of persistence in the latent Markov chain.

Overall, we then considered 16 different possible simulation scenarios. For each of these scenarios we computed a Wald 95 % confidence interval based on the ML estimates and exact estimates for the standard error. In Table 1 we report the empirical coverage and length of the confidence intervals obtained as averages over 1,000 replicates. These values are also averaged over all parameters in each single block. For instance, since the assumed model includes 10 parameters in $\boldsymbol{\phi}$, then the table reports the mean of the coverage and interval length for each of these parameters.

It shall be noted that the actual coverage is very close to 95 % in all cases, with around 50 % of the values below and 50 % above the nominal level on the overall simulation study. We also have checked the results for each single parameter, and the same conclusions apply. It shall also be noted that the inclusion of the covariates does not affect the actual coverage level. The length of the confidence intervals approximately decreases at the rate of $\sqrt{n}$ and slightly slower with T. The use of the covariates tends to increase the average interval length, with a stronger effect of this inclusion for the transition than those for the initial probabilities.

4.2 Application

We considered a large dataset concerning retirement and health among elderly individuals in the United States. Data are collected by the University of Michigan and provided, after some processing, by the RAND Corporation. We used the complete cases of Version I of Health and Retirement Study (HRS) data, downloadable from http://hrsonline.isr.umich.edu. The same data were analyzed in Bartolucci et al. (2013a) by a model based on a different formulation, in which there is an additional effect of the latent process on the response variable with respect to that of the individual covariates.

The analyzed data are referred to a sample of $n = 7,074$ individuals who were asked to express opinions on their health status at $T = 8$ occasions between 1992 and 2006. The outcome is the *self-reported health status*, measured on an ordinal scale divided in five categories: "poor" (coded as 0), "fair" (1), "good" (2), "very good" (3), and "excellent" (4). For every subject we also have information on age, gender (1 for females), race (1 for non-white), and education. The educational level is represented by three categories, and coded with two dummy variables, one for subjects with a college degree and the other for subjects with higher than college degree. Table 2 shows some descriptive statistics about these covariates. A more detailed description of the data can be found in Bartolucci et al. (2013a).

The research question of interest regards the relationship between the self-reported health status and the covariates. This could be useful in order to assess specific needs of different types of individual and to assess the evolution of their health status over time. For this end, we included the covariates only in the latent model component, so as to obtain separate initial and transition probabilities for each configuration of the covariates.

In order to easily illustrate the results we fixed $k = 2$. In this way we are assuming two distinct classes of subjects with a different behavior in terms of tendency to rate their health status. However, if the number of such classes has to be chosen only on the basis of the data, different criteria can be adopted, such as the Bayesian Information Criterion (Schwarz 1978).

Table 2 Distribution of the covariates for the HRS data

Variable	Category	%	Mean	St.dev.
Gender	Male	41.9	–	–
	Female	58.1	–	–
Race	White	82.9	–	–
	Non white	17.1	–	–
Education	Less than college	60.9	–	–
	College or above	39.1	–	–
Age (in 1992)		–	54.8	5.5

Table 3 Estimates of the parameters in $\boldsymbol{\phi}$ for HRS data and corresponding standard errors obtained using the proposed method and the parametric bootstrap method

y	Est.		S.e.		Boot. s.e.	
	$u = 1$	$u = 2$	$u = 1$	$u = 2$	$u = 1$	$u = 2$
0	0.1273	0.0002	0.0023	0.0002	0.0021	0.0001
1	0.3364	0.0058	0.0038	0.0007	0.0028	0.0005
2	0.4551	0.1738	0.0035	0.0038	0.0031	0.0023
3	0.0761	0.5249	0.0028	0.0033	0.0021	0.0031
4	0.0051	0.2954	0.0007	0.0032	0.0006	0.0028

The adopted model has a log-likelihood of −70,865.53 with 29 free parameters, which is considerably higher than that of the independence model, equal to −83,703.21 with 4 parameters, and of the proportional odds model (McCullagh 1980), equal to −80,623.52 with 10 parameters.

With $k = 2$ we obtained the estimates of the conditional response probabilities in $\boldsymbol{\phi}$ which are displayed in Table 3. In this table, we also report the standard errors based on the information matrix obtained by the proposed method and a parametric bootstrap method based on 199 replications. Moreover, in Tables 4, 5, and 6 we report the estimates of parameters in $\boldsymbol{\beta}$ affecting the initial probabilities and the parameters in $\boldsymbol{\gamma}$ affecting the transition probabilities, together with the corresponding standard errors obtained in the two different ways.

The results in Table 3, concerning the conditional distribution of the outcome, allow us to characterize two clearly separate groups: for the first ($u = 1$) a large probability of a "poor", "good" or "fair" health status is observed; subjects in the second group ($u = 2$) tend to have a better opinion, with around 80 % probability for categories "very good" and "excellent". Overall, the second class includes subjects with a better opinion about their health status with respect to the first class.

Concerning the initial probabilities, for $u = 1$ we have an estimate equal to 36 % overall, indicating that at the beginning of the observation period about one third of the subjects are not very satisfied of their health status. This percentage is obtained by averaging the subject-specific initial probabilities corresponding to the parameter estimates in Table 4 for all the sample. These estimates indicate that non-whites are more likely to enter the study in the first latent class given the negative parameter estimate of about −0.96, and with a similar reasoning it can be shown that a higher education is associated with a larger probability of being in the second class. Furthermore, younger subjects are more likely to report a better health status initially. As age increases, satisfaction

Table 4 Estimates of the parameters $\boldsymbol{\beta}$, corresponding standard errors obtained using the proposed method and the parametric bootstrap method, t statistics and related p values

	Est.	S.e.	t test	p value	Boot. s.e.
Intercept	0.5115	0.0696	7.3543	0.0000	0.0641
Gender (F)	−0.0693	0.0643	−1.0770	0.2815	0.0594
Race (non-white)	−0.9554	0.0794	−12.0344	0.0000	0.0779
Education (college)	0.8778	0.0810	10.8370	0.0000	0.0732
Education (>college)	1.6290	0.1000	16.2932	0.0000	0.0983
(Age-50)	−0.0266	0.0071	−3.7246	0.0002	0.0064
$(\text{Age-50})^2/100$	0.0098	0.0537	0.1821	0.8555	0.0498

Table 5 Estimates of the parameters $\boldsymbol{\gamma}_2$, corresponding standard errors obtained using the proposed method and the parametric bootstrap method, t statistics and related p values

	Est.	S.e.	t test	p value	Boot s.e.
Intercept	−4.2840	0.5452	−7.8575	0.0000	0.4229
Gender (F)	−0.6317	0.1943	−3.2515	0.0011	0.1789
Race (non-white)	0.6528	0.1985	3.2880	0.0010	0.1611
Education (college)	−0.1827	0.3054	−0.5981	0.5498	0.2600
Education (>college)	−1.8642	1.0524	−1.7714	0.0765	3.5398
(Age-50)	0.0564	0.0750	0.7520	0.4521	0.0580
$(\text{Age-50})^2/100$	−0.2061	0.2497	−0.8255	0.4091	0.2101

Table 6 Estimates of the parameters $\boldsymbol{\gamma}_1$, corresponding standard errors obtained using the proposed method and the parametric bootstrap method, t statistics and related p values

	Est.	S.e.	t test	p value	Boot. s.e.
Intercept	−2.6025	0.1012	−25.7039	0.0000	0.1044
Gender (F)	−0.3076	0.0671	−4.5834	0.0000	0.0638
Race (non-white)	0.7374	0.0855	8.6220	0.0000	0.0769
Education (college)	−0.3376	0.0826	−4.0894	0.0000	0.0741
Education (>college)	−0.6914	0.0847	−8.1585	0.0000	0.0803
(Age-50)	0.0011	0.0108	0.0986	0.9214	0.0118
$(\text{Age-50})^2/100$	0.0942	0.0418	2.2536	0.0242	0.0421

with one's own health status gets worse, and the effect seems to be linear given that the quadratic effect of age is not significant, with a p value of about 0.86 for the hypothesis that the corresponding parameter is equal to 0.

Concerning the model for the transition probabilities, we report in Table 5 the estimates of the parameters $\boldsymbol{\gamma}_2$, that is, the parameters associated with the transition from $\bar{u} = 1$ to $u = 2$, and in Table 6 the estimates of $\boldsymbol{\gamma}_1$, that is, the parameters associated with the transition from $\bar{u} = 2$ to $u = 1$.

From Tables 5 and 6 we conclude that being non-white decreases the probability of persistence in $\bar{u} = 1$ and $\bar{u} = 2$, that is, non-whites tend to move between latent states more than whites. On the other hand, there seems to be no significant effects of education and age on transitions from $\bar{u} = 1$. As far as a worsening of the opinion regarding one's health status, similar effects are obtained: whites are more likely to persist in the latent state corresponding to a higher satisfaction, a higher degree is associated with persistence as well, and there seems to be a quadratic effect of age increasing the probability of transition. In fact, as age increases, even subjects presently satisfied of their health status tend to move to a less satisfactory situation. A final note regards gender, as a negative log-odds is seen for females both in Tables 5 and 6. We conclude that females are less likely than males to move from one state to another, regardless of the present opinion.

In order to help the interpretation of the results, we also report the average latent transition matrix for the overall sample (Table 7, first panel), which shows a high persistence in the current latent state, with a certain tendency towards a worse health status over time. The number of transitions linked to a worsening in the perceived health status are about seven times those expected in connection with an improvement. In Table 7 we also report the transition matrix estimated for a white woman with a degree above college (second panel) and that estimated for a black man with a lower than college degree (third panel). The results are averaged over all the other covariates. It can be seen that the estimated probabil-

Table 7 Average marginal (first panel) and conditional (second and third panel) latent transition matrices; the second transition matrix is for a white woman with a degree above college; the third is for a black man with a degree below college

	u		u		u	
	1	2	1	2	1	2
$\bar{u}=1$	0.9877	0.0123	0.9985	0.0015	0.9685	0.0348
$\bar{u}=2$	0.0721	0.9279	0.0316	0.9684	0.1681	0.8319

ity of a general worsening of the perceived health status is much larger in the second case than in the first, confirming the effect of the covariate race already commented above.

In conclusion, we note that, for every parameter estimate, the standard error obtained using the information matrix computed by the proposed method is close enough to that obtained by the parametric bootstrap, with few exceptions. One of such exceptions is for some of the parameters in $\boldsymbol{\gamma}_2$, in particular the parameter associated to the highest educational level. However, we checked that this parameter estimate is rather unstable due to the reduced number of subjects with this level of education that are included in the first latent state. We expect that by increasing the number of bootstrap replicates we can obtain more similar results between the proposed method and the bootstrap method. However, this would take a very long computing time in contrast with the immediacy of our method. Therefore, we conclude that through the proposed method we easily obtain reliable standard errors for the parameter estimates even with a very reduced computing time, even when the sample size is relatively large.

Acknowledgments The authors are grateful to an Associate Editor and two Referees for useful comments that helped us to improve the presentation. Francesco Bartolucci acknowledges the financial support from the grant RBFR12SHVV of the Italian Government (FIRB-Futuro in Ricerca-project "Mixture and latent variable models for causal inference and analysis of socio-economic data").

Appendix

Appendix 1: derivatives of the forward-backward recursions

First of all we have that

$$\frac{\partial \log l^{(1)}(u,\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} = \frac{\partial \log \pi(u|\boldsymbol{x}^{(1)})}{\partial\boldsymbol{\theta}} + \frac{\partial \log \phi^{(1)}(\boldsymbol{y}^{(1)}|u,\boldsymbol{w}^{(1)})}{\partial\boldsymbol{\theta}}$$

and

$$\frac{\partial \log l^{(t)}(\bar{u},u,\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} = \frac{\partial \log l^{(t-1)}(\bar{u},\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} + \frac{\partial \log \pi^{(t)}(u|\bar{u},\boldsymbol{x}^{(t)})}{\partial\boldsymbol{\theta}} + \frac{\partial \log \phi^{(t)}(\boldsymbol{y}^{(t)}|u,\boldsymbol{w}^{(t)})}{\partial\boldsymbol{\theta}}.$$

Now considering Eq. (8) we have that

$$\frac{\partial \log l^{(t)}(u,\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} = \sum_{\bar{u}=1}^{k} \frac{l^{(t)}(\bar{u},u,\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{l^{(t)}(u,\tilde{\mathbf{y}}|\tilde{\mathbf{z}})} \frac{\partial \log l^{(t)}(\bar{u},u,\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}},$$

which may be recursively computed for $t=2,\ldots,T$ also taking into account the results in Appendix 2 and that

$$\frac{\partial \log \phi^{(t)}(\boldsymbol{y}|u,\boldsymbol{w})}{\partial\boldsymbol{\alpha}} = \sum_{j=1}^{r} \frac{\partial \log \phi_j^{(t)}(y_j|u,\boldsymbol{w})}{\partial\boldsymbol{\alpha}}.$$

In the end we obtain

$$\frac{\partial \log f(\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} = \sum_{u=1}^{k} \frac{l^{(T)}(u,\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{f(\tilde{\mathbf{y}}|\tilde{\mathbf{z}})} \frac{\partial \log l^{(T)}(u,\tilde{\mathbf{y}}|\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}}.$$

In a similar way we have that

$$\frac{\partial \log m^{(T)}(\tilde{\mathbf{y}}|\bar{u},\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} = 0$$

and

$$\frac{\partial \log m^{(t)}(\tilde{\mathbf{y}}|\bar{u},\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} = \sum_{u=1}^{k} \frac{m^{(t)}(u,\tilde{\mathbf{y}}|\bar{u},\tilde{\mathbf{z}})}{m^{(t)}(\tilde{\mathbf{y}}|\bar{u},\tilde{\mathbf{z}})} \frac{\partial \log m^{(t)}(u,\tilde{\mathbf{y}}|\bar{u},\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}}$$

for $t=2,\ldots,T-1$, where

$$\frac{\partial \log m^{(t)}(u,\tilde{\mathbf{y}}|\bar{u},\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} = \frac{\partial \log m^{(t+1)}(\tilde{\mathbf{y}}|u,\tilde{\mathbf{z}})}{\partial\boldsymbol{\theta}} + \frac{\partial \log \pi^{(t+1)}(u|\bar{u},\boldsymbol{x}^{(t+1)})}{\partial\boldsymbol{\theta}} + \frac{\partial \log \phi^{(t+1)}(\boldsymbol{y}^{(t+1)}|u,\boldsymbol{w}^{(t+1)})}{\partial\boldsymbol{\theta}}.$$

Then these derivatives may be computed by a backward recursion.

Appendix 2: derivatives of the density and probability mass functions

In the case of a canonical GLM parametrization, and considering the general situation of multivariate outcomes, for the measurement component we have

$$\frac{\partial \log \phi_j^{(t)}(y|u,\boldsymbol{w})}{\partial\boldsymbol{\alpha}} = \frac{y-\mu^{(t)}(u,\boldsymbol{w})}{g(\tau)}\boldsymbol{a}_{ju\boldsymbol{w}}^{(t)},$$

$$\frac{\partial^2 \log \phi_j^{(t)}(y|u,\boldsymbol{w})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'} = -V(Y^{(t)}|U^{(t)}=u,\boldsymbol{W}^{(t)}=\boldsymbol{w})\boldsymbol{a}_{ju\boldsymbol{w}}^{(t)}(\boldsymbol{a}_{ju\boldsymbol{w}}^{(t)})',$$

where τ denotes the dispersion parameter and $g(\tau)$ denotes the function involving this parameter in the typical expression for an exponential family distribution (McCullagh and Nelder 1989). In the case of categorical data where a multinomial logit parametrization is adopted, we have

$$\frac{\partial \log \phi_j^{(t)}(y|u,\boldsymbol{w})}{\partial\boldsymbol{\alpha}} = (\boldsymbol{A}_{ju\boldsymbol{w}}^{(t)})'\boldsymbol{G}'_{1c_j}(\boldsymbol{e}_j(y+1)-\boldsymbol{\phi}_j^{(t)}(u,\boldsymbol{w})),$$

where $\boldsymbol{e}_c(y+1)$ is a vector of c zeros with element $y+1$ equal to 1 (because the first category is labelled as 0) and

$$\frac{\partial^2 \log \phi_j^{(t)}(y|u,\boldsymbol{w})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} = (\boldsymbol{A}_{ju\boldsymbol{w}}^{(t)})' \boldsymbol{G}_{1c_j}' \boldsymbol{\Omega}\left(\boldsymbol{\phi}_j^{(t)}(u,\boldsymbol{w})\right) \boldsymbol{G}_{1c_j} \boldsymbol{A}_{ju\boldsymbol{w}}^{(t)},$$

where, for a generic probability vector $\boldsymbol{f}$, we have $\boldsymbol{\Omega}(\boldsymbol{f}) = \mathrm{diag}(\boldsymbol{f}) - \boldsymbol{f}\boldsymbol{f}'$.

Regarding, the other derivatives, we have

$$\frac{\partial \log \pi(u|\boldsymbol{x})}{\partial \boldsymbol{\beta}} = \boldsymbol{B}_{\boldsymbol{x}}' \boldsymbol{G}_{1k}'(\boldsymbol{e}_k(u) - \boldsymbol{\pi}(\boldsymbol{x})),$$

$$\frac{\partial^2 \log \pi(u|\boldsymbol{x})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\boldsymbol{B}_{\boldsymbol{x}}' \boldsymbol{G}_{1k}' \boldsymbol{\Omega}(\boldsymbol{\pi}(\boldsymbol{x})) \boldsymbol{G}_{1k} \boldsymbol{B}_{\boldsymbol{x}},$$

and, finally,

$$\frac{\partial \log \pi^{(t)}(u|\bar{u},\boldsymbol{x})}{\partial \boldsymbol{\gamma}} = \left(\boldsymbol{C}_{\bar{u}\boldsymbol{x}}^{(t)}\right)' \boldsymbol{G}_{\bar{u}k}' \left(\boldsymbol{e}_k(u) - \boldsymbol{\pi}^{(t)}(\bar{u},\boldsymbol{x})\right),$$

$$\frac{\partial^2 \log \pi^{(t)}(u|\bar{u},\boldsymbol{x})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = -\left(\boldsymbol{C}_{\bar{u}\boldsymbol{x}}^{(t)}\right)' \boldsymbol{G}_{\bar{u}k}' \boldsymbol{\Omega}\left(\boldsymbol{\pi}^{(t)}(\bar{u},\boldsymbol{x})\right) \boldsymbol{G}_{\bar{u}k} \boldsymbol{C}_{\bar{u}\boldsymbol{x}}^{(t)},$$

where $\boldsymbol{\pi}(\boldsymbol{x})$ is the column vector of the initial probabilities $\pi(u|\boldsymbol{x})$ and $\boldsymbol{\pi}^{(t)}(\bar{u},\boldsymbol{x})$ is that of the transition probabilities $\pi^{(t)}(u|\bar{u},\boldsymbol{x})$, with $u = 1,\ldots,k$.

References

Agresti, A.: Categorical Data Analysis, 2nd edn. Wiley, Hoboken (2002)

Bartolucci, F., Bacci, S., Pennoni, F.: Longitudinal analysis of self-reported health status by mixture latent autoregressive models. J. R. Stat. Soc. Ser. C, in press (2013a)

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

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

Bartolucci, F., Pandolfi, S.: A new constant memory recursion for hidden Markov models. J. Comput. Biol. **21**, 99–117 (2014)

Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37**, 1554–1563 (1966)

Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. **41**, 164–171 (1970)

Berchtold, A.: Optimization of mixture models: comparison of different strategies. Comput. Stat. **19**, 385–406 (2004)

Cappé, O., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer, New York (2005)

Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)

Farcomeni, A.: Quantile regression for longitudinal data based on latent Markov subject-specific parameters. Stat. Comput. **22**, 141–152 (2012)

Goodman, L.A.: Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika **61**, 215–231 (1974)

Hughes, J.: Computing the observed information in the hidden Markov model using the EM algorithm. Stat. Probab. Lett. **32**, 107–114 (1997)

Khan, R.N.: Statistical Modelling and Analysis of Ion Channel Data Based on Hidden Markov Models and the EM Algorithm. PhD thesis, University of Western Australia, Crawley (2003)

Khreich, W., Granger, E., Miri, A., Sabourin, R.: On the memory complexity of the forward-backward algorithm. Pattern Recognit. Lett. **31**, 91–99 (2010)

Louis, T.A.: Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B **44**, 226–233 (1982)

Lystig, T.C., Hughes, J.: Exact computation of the observed information matrix for hidden Markov models. J. Comput. Gr. Stat. **11**, 678–689 (2002)

McCullagh, P.: Regression models for ordinal data (with discussion). J. R. Stat. Soc. Ser. B **42**, 109–142 (1980)

McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall/CRC Press, London (1989)

McHugh, R.B.: Efficient estimation and local identification in latent class analysis. Psychometrika **21**, 331–347 (1956)

McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, New Jersey (2008)

Oakes, D.: Direct calculation of the information matrix via the EM algorithm. J. R. Stat. Soc. Ser. B **61**, 479–482 (1999)

Orchard, T., Woodbury, M. A.: A missing information principle: theory and applications. In Le Cam, L. M., Neyman, J., and Scott, E. L., (eds.) Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 697–715. Berkeley University of California Press (1972)

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)

Scott, S.L.: Bayesian methods for hidden Markov models: recursive computing in the 21st century. J. Am. Stat. Assoc. **97**, 337–351 (2002)

Turner, R.: Direct maximization of the likelihood of a hidden Markov model. Comput. Stat. Data Anal. **52**, 4147–4160 (2008)

Turner, T.R., Cameron, M.A., Thomson, P.J.: Hidden Markov chains in generalized linear models. Can. J. Stat. **26**, 107–125 (1998)

Welch, L.R.: Hidden Markov models and the Baum–Welch algorithm. IEEE Inf. Theory Soc. Newsl. **53**, 1–13 (2003)

Zucchini, W., MacDonald, I.L.: Hidden Markov Models for Time Series: An Introduction Using R. Chapman & Hall/CRC Press, Boca Raton (2009)