Published in Computational Statistics & Data Analysis
Volume 94, Pages 193-209

Gaussian quadrature approximations in mixed hidden Markov models for longitudinal data: a simulation study

Maria Francesca Marino[a], Marco Alfó[b,*]

[a] *Dipartimento di Economia, Università degli Studi di Perugia*
[b] *Dipartimento di Scienze Statistiche, Sapienza Università di Roma.*

Abstract

Mixed hidden Markov models represent an interesting tool for the analysis of longitudinal data. They allow to account for both time-constant and time-varying sources of unobserved heterogeneity, which are frequent in this kind of studies. Individual-specific latent features, which may be either constant or varying over time, are included in the linear predictor and lead to a general form of dependence between individual measurements. When a parametric (continuous) distribution is associated to time-constant random parameters, the estimation process requires the calculation of (multiple) integrals. These, generally, have not a closed form and should be numerically approximated. Here, the aim is to compare the standard, the adaptive and the pseudo-adaptive Gaussian quadrature approximations by means of a large scale simulation study, where continuous and discrete responses with (conditional) density in the exponential family are considered. Simulation results show that the approximation error is often substantially reduced when the adaptive quadrature rules are considered in place of the standard one. Such an improvement comes at the cost of a higher computational complexity when the fully adaptive scheme is applied. It is shown that, when a sufficient number of repeated measurements per unit is available, the pseudo-adaptive quadrature represents a convenient compromise between quality of results and computational complexity.

Keywords: Hidden Markov models, time-constant and time-varying

*Corresponding author. Marco Alfó, Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 - Rome, ITALY, tel. +390649910763

Email addresses: mariafrancesca.marino@unipg.it (Maria Francesca Marino), marco.alfo@uniroma1.it (Marco Alfó)

random parameters, adaptive Gaussian quadrature, exponential family

1. Introduction

Longitudinal studies entail repeated measurements from a number of units taken over a known, usually finite, time window. In the regression framework, the presence of unobserved individual characteristics, linked e.g. to omitted covariates, leads to extra variability in the marginal distribution of the response and to dependence between measurements from the same individual. Such unobserved heterogeneity can be either time-varying or time-constant, according to a form of *true/spurious contagion* (Heckman, 1981). Because of the former, data variability can be ascribed to the time separation between subsequent measurements: current and future outcomes are directly influenced by the past ones. Because of the latter, differences in the response variable are related to the presence of heterogeneous populations with a different propensity to the event. To account for these sources of extra-variability and dependence, time-constant and time-varying individual random parameters may be added to the model specification. Parametric continuous distributions can be used for both types of random parameters; see Diggle et al. (2002), for references. In a more appealing fashion, the latter can be instead approximated via a discrete latent (hidden) variable with a Markovian structure; the resulting model is referred to as a mixed hidden Markov model (mHMM). It is worth noticing that, when the number of hidden states increases, the discrete Markov process may be able to approximate an AR(1)-type continuous distribution.

If parameter estimates are obtained via a maximum likelihood approach, as it is frequent in the presence of latent variables, the EM algorithm can be employed. Zucchini and MacDonald (2009) and Cappé et al. (2005) give general references, while Bartolucci et al. (2012) discuss a comprehensive overview of applications to longitudinal data. When considering a parametric (continuous) distribution for the time-constant random parameters, ML estimation requires the computation of multiple integrals. Apart from the case when a Gaussian distribution is used for both the response and the random parameters (see e.g. Lagona et al. 2014), these integrals cannot be solved analytically and numerical approximation techniques are a potential solution.

Recently, some proposals have been introduced to deal with such an issue.

Altman (2007) has discussed standard Gaussian quadrature (GQ) in a direct ML perspective and a Monte Carlo EM (MCEM) algorithm; an original variant of the MCEM algorithm has been proposed by Chaubert-Pereira et al. (2010). Maruotti and Rydén (2009) have suggested to leave the distribution of time-constant random parameters unspecified (as in Aitkin, 1999) and to approximate it through a discrete distribution estimated with a nonparametric maximum likelihood (NPML) approach (see Laird 1978; Böhning 1982; Lindsay 1983a,b). For a general review of mixed hidden Markov models the reader is referred to Maruotti (2011).

Although Altman (2007), Maruotti and Rydén (2009) and Lagona et al. (2014) have discussed general random parameter mHMMs, only random intercepts have been considered in empirical applications and simulation studies. Therefore, a first question is whether this class of models can be easily adapted to handle general random parameters. A further question arises when we consider parametric specifications for the distribution of the individual-specific random parameters with time-constant structure. In the context of mixed parameter models, it is generally acknowledged that standard Gaussian quadrature may produce unsatisfactory approximations and poor estimates. Adaptive (Liu and Pierce, 1994; Pinheiro and Bates, 1995) and pseudo-adaptive schemes (Rizopoulos, 2012) have been introduced to improve the quality of results. Within the adaptive quadrature approaches, standard GQ locations, which are symmetric around zero, are centred and scaled at each step (fully adaptive quadrature) or only at the beginning of the optimization algorithm (pseudo-adaptive quadrature) to relocate the main mass of the integrand at zero. This is shown to reduce the approximation error supplied by the GQ technique. In the framework of multilevel models, Rabe-Hesketh et al. (2002, 2005) have proved, via an extensive simulation study, that the adaptive Gaussian quadrature rule outperforms the standard approach, especially when the intraclass correlation is high. Cagnone and Monari (2013) have compared the fully adaptive and the standard Gaussian approximation in the framework of high-dimensional latent variable models; as the dimension increases, the standard Gaussian quadrature turns out to be less appropriate due to the difficulties in reaching convergence in a reasonable number of iterations. In the context of joint models for longitudinal and time to event data, Rizopoulos (2012) has shown that the pseudo-adaptive scheme leads to accurate parameter estimates with a lower number of locations when compared to the standard scheme, thus consistently reducing the computational load.

To our knowledge, this topic has not been adequately investigated in the context of mHMMs; the aim of this paper is at comparing the standard Gaussian quadrature approach discussed by Altman (2007) with the fully adaptive and the pseudo-adaptive quadrature schemes. To assess the quality of these approximations, we have considered, in a large scale simulation study, responses having conditional Gaussian, Poisson and Bernoulli distribution, with varying sample sizes and number of repeated measurements per unit. The plan of the paper follows. In section 2, we introduce the standard mHMM. Sections 3-4 entail the EM algorithm for parameter estimation and the quadrature schemes. Section 5 describes the simulation study and the corresponding results. The last Section contains concluding remarks and outlines future research agenda.

2. Mixed hidden Markov models

As stressed before, these models combine features of hidden Markov and mixed parameter models. In hidden Markov models (see e.g. Zucchini and MacDonald, 2009), the distribution of the observed response is defined conditional on the current hidden state, which represents the realization of a latent process evolving over time according to a Markov structure. In mixed parameter models, see Laird and Ware (1982), the response distribution is specified conditional on individual-specific random parameters that capture latent, time-constant, characteristics. Both models account for marginal dependence between measurements from the same unit.

Before describing mHMMs, some basic notations need to be introduced. Let Y_{it} denote the longitudinal response recorded for unit $i = 1, ..., n$ at occasion $t = 1, ..., T_i$ and let us consider a homogeneous hidden Markov chain $\{S_{it}\}$ taking values in the finite set $\mathscr{S} = \{1, ..., m\}$. In the following, we will refer to measurement occasions that are equally spaced and taken at pre-specified times; for this reason, we will use the generic term *time*. We assume that all individuals share the same initial probability vector $\boldsymbol{\delta} = (\delta_1, ..., \delta_m)$ and the same transition probability matrix $\mathbf{Q} = \{q_{kh}\}$ which is constant over the time. Terms δ_h represent the probability of starting in the h-th state, while q_{kh} represents the probability of moving from the k-th state at time $t-1$ to the h-th one at time t, where $h, k = 1, ..., m, t = 1, ..., T_i$. Let $\mathbf{b}_i$ represent a vector of individual-specific random parameters; a typical choice is to consider Gaussian random parameters $\mathbf{b}_i \sim \text{MVN}(\mathbf{0}, \mathbf{D})$.

mHMMs are based on the following assumptions. The time-constant random

parameters $\mathbf{b}_i$ are independent of the hidden process $\{S_{it}\}$; the distribution of the observed response at a given time is defined conditional on the hidden state occupied at the same time and the individual-specific vector $\mathbf{b}_i$. Conditional on s_{it} and $\mathbf{b}_i$, observations from the same individual are independent (local independence assumption). Based on these hypotheses, the following expression holds

$$f_{y|sb}\big(y_{it} \mid y_{i1:t-1}, s_{i1:t}, \mathbf{b}_i\big) = f_{y|sb}\big(y_{it} \mid s_{it}, \mathbf{b}_i\big),$$

where $y_{i1:t-1}$ represents the history of responses for the i-th individual up to time $t-1$ and $s_{i1:t}$ the sequence of states up to time t. Under the local independence assumption, the joint conditional distribution for the i-th unit longitudinal sequence is defined by

$$f_{y|sb}(\mathbf{y}_i \mid \mathbf{s}_i, \mathbf{b}_i) = \prod_{t=1}^{T_i} f_{y|sb}\big(y_{it} \mid s_{it}, \mathbf{b}_i\big).$$

In the following, we will consider responses with conditional distribution in the exponential family:

$$\big[Y_{it} \mid S_{it} = s_{it}, \mathbf{b}_i\big] \sim EF\left[\theta_{it}(s_{it}, \mathbf{b}_i)\right]. \tag{1}$$

In the h-th state, the canonical parameter is defined by the following regression model

$$\theta_{it}(S_{it} = h, \mathbf{b}_i) = \mathbf{z}'_{it}\left[\boldsymbol{\phi} + \mathbf{b}_i\right] + \mathbf{x}'_{it}\boldsymbol{\beta}_h. \tag{2}$$

Here, $\boldsymbol{\phi}$ represents the marginal effect of covariates in the design vector $\mathbf{z}_{it}$; the $\mathbf{b}_i$'s represent zero mean, time-constant, random departures from $\boldsymbol{\phi}$ due to the effect of omitted covariates. The time-varying effect of such unobserved heterogeneity is captured instead by the state-specific $\boldsymbol{\beta}_h$'s and the corresponding Markov structure. The reasons to adopt such a modelling specification could be explained via a simple example; let us assume the following model holds:

$$\theta_{it} = \mathbf{z}'_{it}\boldsymbol{\phi} + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\boldsymbol{\nu}.$$

If all the potential sources of variation, $[\mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{w}_{it}]$, were observed, the measurements from the same individual would have been independent. However, covariates $\mathbf{w}_{it}$ have not been measured; the effects of such an omission on the remaining model parameters could be either time-constant, summarized by

$\mathbf{b}_i$, or time-varying, summarized by the state-specific parameters $\boldsymbol{\beta}_h$. A side effect of unobserved heterogeneity is that it introduces extra-variability and dependence in the marginal distribution of the observed response. As it is clear from equation (2), the dependence structure postulated by a mHMM is quite general. "Fixed" and "dynamic" random terms (the former, eventually, modulated by the time evolution of $\mathbf{z}_{it}$) allow to account for different sources of unobserved heterogeneity. Obviously, neither of these effects can be simply modelled by adding main effects and interactions with time in the model specification; in fact, dynamics due to unobserved heterogeneity may be strongly non-linear and may not be thought to influence the linear predictor only.
Due to the Markov property, the following expression holds for the density of the individual sequence of states:

$$f_s\left(\mathbf{s}_i; \boldsymbol{\delta}, \mathbf{Q}\right) = f_s\left(s_{i1}; \boldsymbol{\delta}\right) \prod_{t=2}^{T_i} f_s\left(s_{it} \mid s_{it-1}; \mathbf{Q}\right) = \delta_{s_{i1}} \prod_{t=2}^{T_i} q_{s_{it-1}s_{it}}.$$

If we denote by $\boldsymbol{\psi}$ the parameter vector for the longitudinal model, the joint distribution for the individual sequence $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})$ is obtained as

$$f_y(\mathbf{y}_i; \boldsymbol{\psi}, \boldsymbol{\delta}, \mathbf{Q}, \mathbf{D}) = \int \sum_{\mathbf{s}_i} f_{y|sb}\left(\mathbf{y}_i \mid \mathbf{s}_i, \mathbf{b}_i; \boldsymbol{\psi}\right) f_s(\mathbf{s}_i; \boldsymbol{\delta}, \mathbf{Q}) f_b(\mathbf{b}_i; \mathbf{D}) \mathrm{d}\mathbf{b}_i. \quad (3)$$

The multiple integral in (3) does not have, at least generally, a closed form solution; numerical approximation techniques could represent a valid tool. Before detailing such techniques in Section 4, we turn our attention to the algorithm for ML estimation.

3. Maximum likelihood estimation

As it is frequent with latent variable models, also in the case of mHMMs the EM algorithm (see Dempster et al. 1977) may be employed for parameter estimation. Further alternatives are available: the direct maximization of the likelihood function (Altman, 2007; Langrock et al., 2014), the MCEM algorithm described, in a very simple development, by Altman (2007) and in a quite more elaborated form by Chaubert-Pereira et al. (2010). However, in our opinion, the EM algorithm represents a conventional choice, it is computationally simple and widely adopted by non-statisticians also. For these

reasons, we have decided to focus on it.
We denote by $\mathbf{\Phi} = (\boldsymbol{\psi}, \boldsymbol{\delta}, \mathbf{Q}, \mathbf{D})$ the full set of model parameters; let $u_{it}(h) = \mathbb{I}[S_{it} = h]$ be the indicator variable for the i-th individual in the h-th state at time t and $u_{it}(k,h) = \mathbb{I}[S_{it-1} = k, S_{it} = h]$ be equal to 1 if the i-th individual moves from the k-th state at time $t-1$ to the h-th one at time t. The complete data log-likelihood follows:

$$\ell_c(\mathbf{\Phi}) = \sum_{i=1}^{n} \left\{ \sum_{h=1}^{m} u_{i1}(h) \log(\delta_h) + \sum_{t=2}^{T_i} \sum_{h=1}^{m} \sum_{k=1}^{m} u_{it}(k,h) \log(q_{kh}) + \sum_{t=1}^{T_i} \sum_{h=1}^{m} u_{it}(h) \log\left[f_{y|sb}(y_{it} \mid S_{it} = h, \mathbf{b}_i; \boldsymbol{\psi})\right] + \log\left[f_b(\mathbf{b}_i; \mathbf{D})\right] \right\}. \quad (4)$$

The expression within brackets represents the individual contribution to the complete data log-likelihood; in the following, it will be denoted by $\ell_c^{(i)}, i = 1, ...n$. For ease of notation, we will suppress the dependence of density functions on model parameters.
To simplify the estimation procedure, we introduce the forward and the backward variables (Baum et al., 1970); here, the basic definition should be modified to account for the presence of the individual-specific random parameters $\mathbf{b}_i$. The forward variables represent, for a generic individual, the joint density of the longitudinal measurements up to time t and of ending in the h-th state, conditional on the individual-specific vector $\mathbf{b}_i$:

$$a_{it}(h, \mathbf{b}_i) = f\left(y_{i1:t}, S_{it} = h \mid \mathbf{b}_i\right). \quad (5)$$

By modifying the arguments in Baum et al. (1970), forward variables $a_{it}(h, \mathbf{b}_i)$ can be recursively computed as

$$a_{i1}(h, \mathbf{b}_i) = \delta_h f_{y|sb}(y_{i1} \mid S_{i1} = h, \mathbf{b}_i),$$
$$a_{it}(h, \mathbf{b}_i) = \sum_{k=1}^{m} a_{it-1}(k, \mathbf{b}_i) q_{kh} f_{y|sb}(y_{it} \mid S_{it} = h, \mathbf{b}_i).$$

The backward variables represent the probability of the longitudinal sequence from time $t+1$ to the last observation, conditional on being in the h-th state at time t and the random parameter vector $\mathbf{b}_i$:

$$b_{it}(h, \mathbf{b}_i) = f(y_{it+1:T_i} \mid S_{it} = h, \mathbf{b}_i). \quad (6)$$

Also backward variables can be derived recursively:

$$b_{iT_i}(h, \mathbf{b}_i) = 1,$$
$$b_{it-1}(k, \mathbf{b}_i) = \sum_{h=1}^{m} b_{it}(h, \mathbf{b}_i) q_{kh} f_{y|sb}(y_{it} \mid S_{it} = h, \mathbf{b}_i),$$

These recursions greatly simplify the structure of the estimation algorithm; for further details, see the seminal paper by Baum et al. (1970), the reference monograph by Zucchini and MacDonald (2009) and the monograph by Bartolucci et al. (2012) that gives a specific overview in the longitudinal data context.

3.1. The E-step

The E-step involves calculating the expected value of the complete data log-likelihood (4) given the observed data and the current parameter estimates:

$$Q(\mathbf{\Phi} \mid \mathbf{\Phi}^{(r)}) = \sum_{i=1}^{n} \int \sum_{\mathbf{s}_i} \ell_c^{(i)}(\mathbf{\Phi}) f_{sb|y}(\mathbf{s}_i, \mathbf{b}_i \mid \mathbf{y}_i, \mathbf{\Phi}^{(r)}) \mathrm{d}\mathbf{b}_i.$$

Following Rijmen et al. (2008), we may notice that the crucial quantities that are needed in the M-step are the posterior densities of marginal and pairwise consecutive state probabilities. By adapting this result to the mHMM framework, we get

$$\begin{aligned} Q(\mathbf{\Phi} \mid \mathbf{\Phi}^{(r)}) = \sum_{i=1}^{n} \Bigg\{ & \sum_{h=1}^{m} \hat{u}_{i1}(h)^{(r+1)} \log(\delta_h) + \sum_{t=2}^{T_i} \sum_{h,k=1}^{m} \hat{u}_{it}(k,h)^{(r+1)} \log(q_{kh}) + \\ & + \sum_{t=1}^{T_i} \sum_{h=1}^{m} \int \Big[\hat{u}_{it}(h)^{(r+1)} \log \left[f_{y|sb}\left(y_{it} \mid S_{it} = h, \mathbf{b}_i\right) \right] \times \\ & \times f_{b|sy}\big(\mathbf{b}_i \mid S_{it} = h, \mathbf{y}_i; \mathbf{\Phi}^{(r)}\big) \Big] \mathrm{d}\mathbf{b}_i + \int \log \left[f_b(\mathbf{b}_i) \right] f_{b|y}\big(\mathbf{b}_i \mid \mathbf{y}_i; \mathbf{\Phi}^{(r)}\big) \mathrm{d}\mathbf{b}_i \Bigg\}, \end{aligned} \tag{7}$$

where the terms $\hat{u}_{it}(h)^{(r+1)}$ and $\hat{u}_{it}(k,h)^{(r+1)}$ denote the posterior expectations of the indicator variables in equation (4), given the current parameter estimates, $\mathbf{\Phi}^{(r)}$, and the observed data, $\mathbf{y}_i$. Suppressing the dependence on

the iteration index and doing a little algebra, posterior probabilities can be computed as

$$\hat{u}_{it}(h) = \frac{\int a_{it}(h, \mathbf{b}_i) b_{it}(h, \mathbf{b}_i) f_b(\mathbf{b}_i) \mathrm{d}\mathbf{b}_i}{\int \sum_h a_{it}(h, \mathbf{b}_i) b_{it}(h, \mathbf{b}_i) f_b(\mathbf{b}_i) \mathrm{d}\mathbf{b}_i}, \tag{8}$$

and

$$\hat{u}_{it}(k, h) = \frac{\int a_{it-1}(k, \mathbf{b}_i) q_{kh} f_{y|sb}\left(y_{it} \mid S_{it} = h, \mathbf{b}_i\right) b_{it}(h, \mathbf{b}_i) f_b(\mathbf{b}_i) \mathrm{d}\mathbf{b}_i}{\int \sum_{hk} a_{it-1}(k, \mathbf{b}_i) q_{kh} f_{y|sb}\left(y_{it} \mid S_{it} = h, \mathbf{b}_i\right) b_{it}(h, \mathbf{b}_i) f_b(\mathbf{b}_i) \mathrm{d}\mathbf{b}_i}. \tag{9}$$

These quantities require the calculation of multiple integrals which, generally, do not have a closed form solution and have to be numerically approximated. The point is further discussed in Section 4.

3.2. The M-step

The M-step of the algorithm consists in maximizing the expected value of the complete data log-likelihood in equation (7) with respect to $\boldsymbol{\Phi} = (\boldsymbol{\psi}, \boldsymbol{\delta}, \mathbf{Q}, \mathbf{D})$. Due to the local independence assumption and to the separability of the parameter spaces, the maximization can be partitioned into different sub-problems. The maximization can be performed sequentially with respect to the Markov chain parameters, the longitudinal model parameters and the covariance matrix of the random parameters. At the generic iteration of the algorithm, closed form solutions are available for the initial and the transition probabilities:

$$\hat{\delta}_h = \frac{\sum_{i=1}^n \hat{u}_{i1}(h)}{n}, \quad \hat{q}_{kh} = \frac{\sum_{i=1}^n \sum_{t=1}^{T_i} \hat{u}_{it}(k, h)}{\sum_{i=1}^n \sum_{t=1}^{T_i} \sum_{k=1}^m \hat{u}_{it}(k, h)}. \tag{10}$$

Estimation of the longitudinal model parameters reduces to finding the zeros of the expected score function, calculated with respect to the posterior distribution of the hidden states and the random parameters:

$$\sum_{i=1}^n \sum_{t=1}^{T_i} \sum_{h=1}^m \hat{u}_{it}(h) \int \mathbb{S}_{it}\left(\boldsymbol{\psi} \mid S_{it} = h, \mathbf{b}_i\right) f_{b|sy}\big(\mathbf{b}_i \mid S_{it} = h, \mathbf{y}_i\big) \, \mathrm{d}\mathbf{b}_i, \tag{11}$$

where $\mathbb{S}_{it}\left(\boldsymbol{\psi} \mid h, \mathbf{b}_i; y_{it}\right)$ is the individual contribution to the complete data score function for a generic unit being in the h-th state at time t.

We have adopted a restricted maximum likelihood approach (Patterson and Thompson, 1971) to estimate the covariance matrix $\mathbf{D}$; REML allows to correct the bias of standard ML estimator and leads to the following expression:

$$\widehat{\mathbf{D}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{b}_i\mathbf{b}_i^{\mathsf{T}} \mid \mathbf{y}_i\right]. \tag{12}$$

As it is clear, parameter estimation needs the integrals in equations (8)-(9) and (11)-(12) to be numerically approximated.

4. Integration via Gaussian quadrature

As we have pointed out before, the EM algorithm needs the calculation of multiple integrals which, often, do not have a closed form solution. In this context, techniques based on Gaussian quadrature approximations are a main choice due to the long-standing use in the context of mixed parameter models. The general idea is to rephrase the multivariate integral as the product of several univariate integrals; each of them may be approximated through a weighted sum over a pre-specified number of known quadrature abscissas with known weights.
Let us focus on solving likelihood equations associated to the score in equation (11), since calculation of the integrals in equations (8), (9) and (12) proceeds following similar arguments. To simplify the notation, we will consider only the expression appearing within the integral. Let $\mathbf{D} = \mathbf{\Gamma}\mathbf{\Gamma}'$ be the Cholesky decomposition of the random parameter covariance matrix and let q be the dimension of $\mathbf{b}_i$. Using a standard Gaussian quadrature (GQ) approximation, the following expression for the expected score function holds:

$$\begin{aligned}
&\int \mathbb{S}_{it}(\boldsymbol{\psi} \mid S_{it} = h, \mathbf{b}_i) f_{b|sy}(\mathbf{b}_i \mid S_{it} = h, \mathbf{y}_i)\mathrm{d}\mathbf{b}_i \simeq \\
&\simeq \sqrt{2^q}|\mathbf{\Gamma}| \sum_{g_1 \dots g_q} \mathbb{S}_{it}(\boldsymbol{\psi} \mid S_{it} = h, \mathbf{b}^{\star}_{g_1 \dots g_q}) f_{b|sy}(\mathbf{b}^{\star}_{g_1 \dots g_q} \mid S_{it} = h, \mathbf{y}_i) \mathrm{e}^{\|\mathbf{b}_{g_1 \dots g_q}\|^2} w_g,
\end{aligned} \tag{13}$$

where $\mathbf{b}_{g_1 \dots g_q} = \left(b_{g_1}, ..., b_{g_q}\right)$ represents a q-tuple defined by the Cartesian product of standard quadrature points, for $g_l \in \{1, \dots, G\}, l \in \{1, \dots, q\}$, and $w_g = \prod_{l=1}^{q} w_{g_l}$ denotes the product of the corresponding weights. The rescaled abscissas $\mathbf{b}^{\star}_{g_1 \dots g_q} = \sqrt{2}\mathbf{\Gamma}\mathbf{b}_{g_1 \dots g_q}$ allow to rewrite the integral with respect to orthogonal components.

Even if the quality of the GQ approximation is known to increase with the number of quadrature points, the corresponding locations are symmetric around zero and it turns out to be satisfactory only if the integrand has its own peak near zero. When this condition is not satisfied, GQ leads to poor approximations, even for a large number of locations, see e.g. Rizopoulos (2012).

To improve the quality of the estimates, we may consider an adaptive Gaussian quadrature scheme (aGQ). It involves the calculation, at each step of the algorithm, of the posterior modes and curvatures of the random parameter distribution. These are used to shift and scale GQ locations, placing the peak of the integrand function at zero. This gives a Gaussian density having the same logarithmic derivatives up to second order (at the mode) as the integrand function, see Liu and Pierce (1994). Posterior modes are found by solving the problem

$$\widehat{\mathbf{b}}_i = \arg\max_{\mathbf{b}_i} \left\{ \log \sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i) \right\}, \tag{14}$$

while posterior curvatures are locally approximated by the inverse of the negative Hessian matrix, evaluated at the mode:

$$\widehat{\mathbf{H}}_i = \left[-\frac{\partial^2 \log \sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^{\mathsf{T}}} \right]_{\mathbf{b}_i = \widehat{\mathbf{b}}_i}. \tag{15}$$

Denoting by $\boldsymbol{\Gamma}_i$ the Cholesky factorization of $\widehat{\mathbf{H}}_i^{-1}$, we approximate the integral in equation (11) through the following weighted sum:

$$\begin{aligned} &\int \mathbb{S}_{it}(\boldsymbol{\psi} \mid S_{it} = h, \mathbf{b}_i) f_{b|sy}(\mathbf{b}_i \mid S_{it} = h, \mathbf{y}_i) \mathrm{d}\mathbf{b}_i \simeq \\ &\simeq \sqrt{2^q} |\boldsymbol{\Gamma}_i| \sum_{g_1 \ldots g_q} \mathbb{S}_{it}(\boldsymbol{\psi} \mid S_{it} = h, \widehat{\mathbf{b}}^{\star}_{g_1 \ldots g_q}) f_{b|sy}(\widehat{\mathbf{b}}^{\star}_{g_1 \ldots g_q} \mid S_{it} = h, \mathbf{y}_i) \mathrm{e}^{\|\mathbf{b}_{g_1 \ldots g_q}\|^2} w_g, \end{aligned} \tag{16}$$

where $\widehat{\mathbf{b}}^{\star}_{g_1 \ldots g_q} = \widehat{\mathbf{b}}_i + \sqrt{2} \boldsymbol{\Gamma}_i \mathbf{b}_{g_1 \ldots g_q}$.

As noticed by Pinheiro and Bates (1995), ordinary quadrature is a deterministic version of Monte Carlo integration (MC), while adaptive quadrature is a deterministic version of importance sampling (IS). Generally, IS turns out to be computationally more efficient when compared to standard MC

integration; similarly, when compared to the standard Gaussian quadrature rule, aGQ usually needs a reduced number of locations, see e.g. Rizopoulos (2012). Therefore, we may expect that, also in the mHMM context, adaptive quadrature routines produce more precise parameter estimates with a lower number of quadrature points and limited convergence issues. We should however bear in mind that, in the mHMM context, computation of the random parameter posterior distribution may represent a challenging task. Such a distribution is obtained by summing over all possible states of the hidden Markov chain; while the sum can be efficiently solved by using an appropriate matrix product, the step still remains computationally demanding. In fact, we may encounter flat regions in the likelihood surface that do not help in deriving accurate estimates for the posterior modes and curvatures. To skip the problem of working with the posterior distribution of $\mathbf{b}_i$, Altman (2007) proposed to use a MCEM algorithm for parameter estimation. However, as pointed out by Hartzel et al. (2001), an important issue with standard Monte Carlo EM procedures concerns the number of points that have to be sampled for the integral to be adequately approximated. Altman (2007) found that a good value for the number of samples to be drawn is $B = 5000$, while Chaubert-Pereira et al. (2010) used an adaptive approach by sampling a progressively increasing number of points as a function of the current iteration number. In both cases, clearly, the algorithm for parameter estimation is quite demanding.

In the context of joint models for longitudinal and time to event data, Rizopoulos (2012) introduces a pseudo-adaptive quadrature scheme to reduce the computational load of the fully adaptive approach. When the number of repeated measurements increases, it is not necessary to update posterior modes and curvatures of the random parameters at each step of the algorithm. These quantities are estimated only at the beginning of the optimization routine and are used to centre and scale Gauss-Hermite locations which, therefore, remain constant throughout the iterations.

In the present context, we may define a similar approach. Let us re-write the posterior density of the random parameters (on the log scale) as

$$\log \left[f_{b|y}(\mathbf{b}_i \mid \mathbf{y}_i) \right] \propto \log \left[f_b(\mathbf{b}_i) \right] + \log \left\{ \sum_{\mathbf{s}_i} \left[f_{y|sb}(\mathbf{y}_i \mid \mathbf{s}_i, \mathbf{b}_i) f_s(\mathbf{s}_i) \right] \right\}.$$

As T_i increases, $\log \sum_{\mathbf{s}_i} \left[f_{y|sb}(\mathbf{y}_i \mid \mathbf{s}_i, \mathbf{b}_i) f_s(\mathbf{s}_i) \right]$ is the leading term. Following these arguments, standard quadrature locations are centred and scaled only

once via posterior modes and curvatures computed at the first step of the algorithm by marginalizing the mHMM with respect to the hidden Markov process. Our practical experience with an algorithm based on the multiple summation $\sum_{s_{i1}} \cdots \sum_{s_{iT_i}}$ is that such a complex structure often leads to issues of numerical instability. Therefore, a good strategy to derive approximate posterior modes and curvatures starts from fitting a (non linear) mixed model based on $f_{y|b}(\mathbf{y}_i \mid \mathbf{b}_i)$ for which, in turn, we assume the same parametric form as $f_{y|sb}(\mathbf{y}_i \mid \mathbf{s}_i, \mathbf{b}_i)$.
As a result, Gauss-Hermite locations are centred and scaled only once, with respect to the random parameter posterior modes, $\bar{\mathbf{b}}_i$, and curvatures, $\bar{\mathbf{H}}_i^{-1}$, obtained by estimating either a mHMM (PaGQ$_1$) or a mixed model (PaGQ$_2$) with canonical parameter defined by

$$\theta_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{b}_i.$$

Denoting by $\bar{\boldsymbol{\Gamma}}_i$ the Cholesky factor of the posterior curvature $\bar{\mathbf{H}}_i^{-1}$ (evaluated at the posterior mode), the following approximation holds:

$$\begin{aligned}
&\int \mathbb{S}_{it}(\boldsymbol{\psi} \mid S_{it} = h, \mathbf{b}_i) f_{b|sy}(\mathbf{b}_i \mid S_{it} = h, \mathbf{y}_i) \mathrm{d}\mathbf{b}_i \simeq \\
&\simeq \sqrt{2^q} |\bar{\boldsymbol{\Gamma}}_i| \sum_{g_1 \ldots g_q} \mathbb{S}_{it}(\boldsymbol{\psi} \mid S_{it} = h, \bar{\mathbf{b}}^{\star}_{g_1 \ldots g_q}) f_{b|sy}(\bar{\mathbf{b}}^{\star}_{g_1 \ldots g_q} \mid S_{it} = h, \mathbf{y}_i) e^{\|\mathbf{b}_{g_1 \ldots g_q}\|^2} w_g,
\end{aligned} \tag{17}$$

where, $\bar{\mathbf{b}}^{\star}_{g_1 \ldots g_q} = \bar{\mathbf{b}}_i + \sqrt{2}\bar{\boldsymbol{\Gamma}}_i \mathbf{b}_{g_1 \ldots g_q}, g_l = 1, \ldots, G, l = 1, \ldots, q$.
Gaussian quadrature approximations are based on a parametric specification of the random parameter distribution which could be less efficient and less appropriate to describe multimodal, strongly asymmetric, distributions. To avoid potential bias due to misspecification of the random parameter distribution, Maruotti and Rydén (2009) estimate locations and masses of the mixing distribution by using a NPML approach. However, when we move from random intercept to random parameter models, the NPML approach is known to face some difficulties in recovering the *true* covariance structure. Neuhaus et al. (2013) noticed that, even when the random parameter distribution is misspecified, the bias in the parameter estimates is often limited; however, caution is needed since the findings heavily depend on the model structure.
In the next section, results obtained from a large scale simulation study are presented to evaluate the quality of the approximation obtained under the different quadrature schemes we have discussed so far.

5. Simulation study

To compare the performance of the proposed approximations, we have conducted the following simulation study.

5.1. Simulation design

We have randomly drawn $B = 500$ samples from a mHMM considering conditional Gaussian, Poisson and Bernoulli distributions. Four possible experimental scenarios have been considered, varying the sample size, $n = 250, 500$, and the number of (equally spaced) times, $T = 6, 10$. Since missingness is a common problem in longitudinal data, we have considered a missing at random drop-out process. For each unit, the last measurement time (T_i) has been drawn from a discrete distribution defined over the support $\{1, ..., T\}$, with $\Pr(T_i = T) = 0.8$ and $\Pr(T_i = t) = 0.2/(T-1)$, $\forall t \in \{1, \ldots, T-1\}$. In all cases, we have considered two hidden states ($m = 2$) and the following regression model to describe the canonical parameter in the h-th state, $h = 1, \ldots, m$:

$$\theta_{it}(S_{it} = h, \mathbf{b}_i) = [\mathbf{1}, \mathbf{z}_{it}]' \boldsymbol{\phi} + \mathbf{x}'_{it}\boldsymbol{\beta}_h + \mathbf{z}'_{it}\mathbf{b}_i, \quad t = 1, ..., T_i.$$

The covariates associated with the time-constant random parameters are $\mathbf{z}_{it} = [z_{it1}, z_{it2}]$, with $z_{it1} \sim \text{Unif}(-1, 0)$ and $z_{it2} = z_{i2} \sim \text{N}(0, 0.5)$; the covariates associated with the state-specific effects, $\mathbf{x}_{it}$, have been drawn from a bivariate Gaussian distribution, $\text{MVN}(\mathbf{1}, 0.5\mathbf{I}_2)$.
In the Gaussian case, a constant error variance, $\sigma^2_{\text{err}} = 1$, has been considered; fixed parameters have been set to $\boldsymbol{\phi} = (2, 1.4, -0.6)$, while state-specific parameters have been fixed to $\boldsymbol{\beta}_1 = (1.6, 1.4)$ and $\boldsymbol{\beta}_2 = (0.9, 0.5)$.
In the Poisson case, the following set of model parameters has been considered: $\boldsymbol{\phi} = (-0.8, -0.6, 0.4)$; $\boldsymbol{\beta}_1 = (0.1, 0.2)$ and $\boldsymbol{\beta}_2 = (0.5, 0.7)$. Finally, in the Bernoulli case, we have set $\boldsymbol{\phi} = (0.5, -0.7, -0.6)$, $\boldsymbol{\beta}_1 = (1.5, -0.4)$ and $\boldsymbol{\beta}_2 = (0.5, -1.4)$. Both in the continuous and the discrete data scenarios, random parameters have been drawn from a multivariate Gaussian distribution, with upper triangular covariance matrix equal to $(1, 0.5, 2)$. Initial and transition probabilities for the Markov chain have been fixed to

$$\boldsymbol{\delta} = (0.2, 0.8) \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} 0.7 & 0.3 \\ 0.1 & 0.9 \end{pmatrix}. \tag{18}$$

Therefore, with time passing by, the second state of the chain is more and more frequent, and the first one becomes almost empty. This choice has

been done in order to study the precision of parameter estimates associated to hidden states with low probability. For each of the quadrature schemes, we have considered $G = 7, 9, 11$ quadrature points to approximate the integrals but we did not find any relevant change. Therefore, we have chosen $G = 7$ to guarantee a good balance between quality of the approximation and computational complexity. In the following subsections simulation results are discussed.

5.2. Simulation results - the Gaussian case

Tables 1 and 2 report the bias and the standard deviation of parameter estimates calculated over $B = 500$ samples. We must remind that, in this case, mHMM reduces to a simple HMM with a more complex covariance structure. In fact, we have

$$\left.\begin{array}{l}[\mathbf{Y}_i \mid \mathbf{s}_i, \mathbf{b}_i] \sim \text{MVN}(\boldsymbol{\theta}_i(\mathbf{s}_i, \mathbf{b}_i), \sigma^2\mathbf{I}) \\ [\mathbf{b}_i] \sim \text{MVN}(\mathbf{0}, \mathbf{D})\end{array}\right\} \Rightarrow [\mathbf{Y}_i \mid \mathbf{s}_i] \sim \text{MVN}(\boldsymbol{\theta}_i(\mathbf{s}_i), \sigma^2\mathbf{I}+\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i').$$

Therefore, all the approaches refer to the same (closed form) solution and we do not expect relevant differences in the precision of parameter estimates and in the computational times. We show the results obtained from the standard (GQ), the fully adaptive (aGQ) and the pseudo-adaptive quadrature approach starting from a mHMM (PaGQ_1) with $G = 7$ locations.
As we may notice, in all the scenarios, the three approaches return quite comparable results, both in terms of bias and standard deviation. A higher variability can be observed for the variance components when compared to the other parameter estimates, but this variability reduces with increasing T and n. When a limited number of repeated measurements is available (i.e. $T = 6$), PaGQ_1 produces parameter estimates that generally have a reduced bias than that of the standard GQ approach, but a higher variability. However, as the number of times increases ($T = 10$), we may observe that, even in this toy scenario, PaGQ_1 ensures a slight reduction of the approximation error supplied by the standard Gaussian quadrature approach. Thus, the results obtained through the proposed quadrature approaches are coherent with the simulation design. Only slight, mainly numerical, differences, can be appreciated when comparing the different methods. Therefore, we may proceed to consider cases where the integrals defining the log-likelihood need to be numerically evaluated, in order to understand the relative behaviour of the analysed approaches.

Table 1: Simulation study. Mean parameter estimates for the Gaussian mHMM with $n = 250, m = 2$ and $G = 7$

T = 6		GQ		aGQ		PaGQ_1	
		BIAS	SD	BIAS	SD	BIAS	SD
δ_1	0.2	0.024	0.095	0.023	0.090	0.012	0.095
δ_2	0.8	-0.024	0.095	-0.023	0.090	-0.012	0.095
q_{11}	0.7	-0.051	0.181	-0.043	0.158	-0.041	0.185
q_{12}	0.3	0.051	0.181	0.043	0.158	0.041	0.185
q_{21}	0.1	0.152	0.280	0.131	0.260	0.126	0.274
q_{22}	0.9	-0.152	0.280	-0.131	0.260	-0.126	0.274
ϕ_0	2	-0.003	0.122	0.000	0.123	-0.004	0.124
ϕ_1	1.4	-0.003	0.137	-0.002	0.137	0.011	0.144
ϕ_2	-0.6	0.000	0.184	-0.003	0.160	0.009	0.193
β_{11}	1.6	-0.027	0.127	-0.019	0.121	-0.026	0.125
β_{21}	1.4	-0.031	0.137	-0.022	0.133	-0.034	0.137
β_{12}	0.9	0.127	0.272	0.108	0.263	0.093	0.282
β_{22}	0.5	0.172	0.331	0.148	0.313	0.131	0.332
σ_{11}	1	0.056	0.242	0.033	0.237	0.044	0.282
σ_{12}	0.5	-0.025	0.237	-0.024	0.233	0.014	0.279
σ_{22}	2	0.060	0.541	0.013	0.506	-0.126	0.735
T = 10		GQ		aGQ		PaGQ_1	
		BIAS	SD	BIAS	SD	BIAS	SD
δ_1	0.2	0.027	0.087	0.022	0.081	0.019	0.087
δ_2	0.8	-0.027	0.087	-0.022	0.081	-0.019	0.087
q_{11}	0.7	-0.063	0.177	-0.043	0.139	-0.058	0.165
q_{12}	0.3	0.063	0.177	0.043	0.139	0.058	0.165
q_{21}	0.1	0.119	0.251	0.097	0.233	0.106	0.241
q_{22}	0.9	-0.119	0.251	-0.097	0.233	-0.106	0.241
ϕ_0	2	0.001	0.091	0.002	0.090	-0.002	0.090
ϕ_1	1.4	-0.007	0.109	-0.009	0.110	0.003	0.113
ϕ_2	-0.6	-0.003	0.183	-0.004	0.149	-0.002	0.176
β_{11}	1.6	-0.030	0.103	-0.020	0.096	-0.028	0.100
β_{21}	1.4	-0.034	0.120	-0.021	0.104	-0.031	0.110
β_{12}	0.9	0.112	0.245	0.087	0.219	0.098	0.234
β_{22}	0.5	0.144	0.309	0.105	0.271	0.117	0.288
σ_{11}	1	0.041	0.192	0.015	0.184	0.016	0.190
σ_{12}	0.5	-0.010	0.207	-0.007	0.201	0.009	0.214
σ_{22}	2	0.075	0.459	0.018	0.427	0.021	0.452

Table 2: Simulation study. Mean parameter estimates for the Gaussian mHMM with $n = 500, m = 2$ and $G = 7$

T = 6		GQ		aGQ		PaGQ_1	
		BIAS	SD	BIAS	SD	BIAS	SD
δ_1	0.2	0.021	0.088	0.018	0.079	0.005	0.082
δ_2	0.8	-0.021	0.088	-0.018	0.079	-0.005	0.082
q_{11}	0.7	-0.075	0.185	-0.052	0.147	-0.057	0.182
q_{12}	0.3	0.075	0.185	0.052	0.147	0.057	0.182
q_{21}	0.1	0.148	0.264	0.121	0.246	0.122	0.256
q_{22}	0.9	-0.148	0.264	-0.121	0.246	-0.122	0.256
ϕ_0	2	0.002	0.084	0.003	0.083	-0.001	0.084
ϕ_1	1.4	-0.005	0.093	-0.007	0.093	0.006	0.101
ϕ_2	-0.6	-0.005	0.130	-0.005	0.117	-0.003	0.144
β_{11}	1.6	-0.031	0.094	-0.019	0.086	-0.031	0.089
β_{21}	1.4	-0.043	0.111	-0.029	0.105	-0.040	0.105
β_{12}	0.9	0.146	0.253	0.116	0.227	0.110	0.252
β_{22}	0.5	0.178	0.313	0.142	0.285	0.132	0.304
σ_{11}	1	0.040	0.186	0.014	0.178	0.025	0.217
σ_{12}	0.5	-0.011	0.169	-0.011	0.163	0.025	0.208
σ_{22}	2	0.118	0.398	0.064	0.376	-0.066	0.599
T = 10		GQ		aGQ		PaGQ_1	
		BIAS	SD	BIAS	SD	BIAS	SD
δ_1	0.2	0.023	0.084	0.016	0.077	0.016	0.082
δ_2	0.8	-0.023	0.084	-0.016	0.077	-0.016	0.082
q_{11}	0.7	-0.063	0.181	-0.044	0.143	-0.062	0.178
q_{12}	0.3	0.063	0.181	0.044	0.143	0.062	0.178
q_{21}	0.1	0.126	0.256	0.111	0.246	0.117	0.251
q_{22}	0.9	-0.126	0.256	-0.111	0.246	-0.117	0.251
ϕ_0	2	0.006	0.065	0.008	0.065	0.004	0.065
ϕ_1	1.4	-0.004	0.080	-0.006	0.080	0.007	0.086
ϕ_2	-0.6	0.000	0.123	-0.003	0.101	0.011	0.127
β_{11}	1.6	-0.028	0.086	-0.021	0.079	-0.028	0.082
β_{21}	1.4	-0.044	0.104	-0.035	0.099	-0.043	0.098
β_{12}	0.9	0.114	0.228	0.091	0.213	0.099	0.224
β_{22}	0.5	0.158	0.293	0.127	0.268	0.141	0.290
σ_{11}	1	0.047	0.141	0.021	0.137	0.022	0.140
σ_{12}	0.5	-0.023	0.133	-0.017	0.133	0.003	0.148
σ_{22}	2	0.057	0.341	0.009	0.323	0.009	0.345

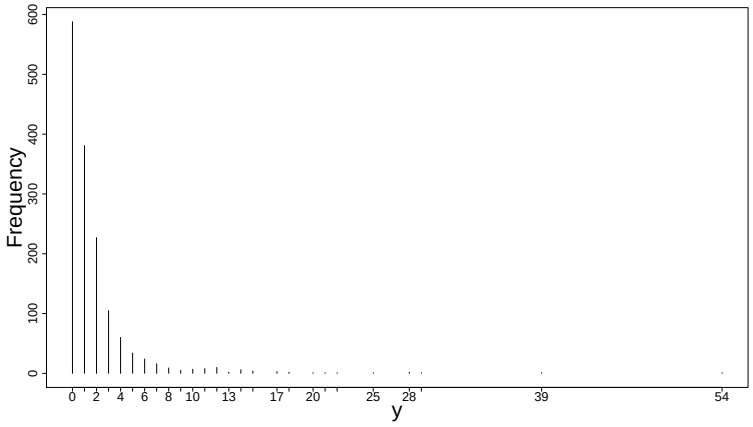
5.3. Simulation results - the Poisson case

Tables 3 and 4 report the results for the Poisson scenario. In this case, the posterior distribution of the random parameters takes a further step away from the standard Gaussian. Figure 1 reports the response variable distribution of a generic simulation sample to give an idea of how far from the Gaussianity we are.

By looking at simulation results, we may observe that all the considered approximation methods turn out to be less accurate than in the previous case. If we compare the standard and the adaptive Gaussian quadrature approaches, we may observe a reduced bias (particularly for the variance components) and a reduced variability of parameter estimates (in some cases this variability is more than halved, e.g. for the estimates of fixed parameters in the longitudinal data model) when the aGQ approach is employed. The gap between the two methods seems to increase as n and T increase.

Focusing on the pseudo-adaptive approaches, as expected, the quality of results is a bit lower when compared to the aGQ approximation, even if the distance seems to reduce with increasing sample size and number of measurement occasions. On the other hand, if we consider GQ as the competitor, both PaGQ_1 and PaGQ_2 turn out to be globally more efficient. As it has been said before, posterior modes and curvatures for PaGQ_1 come from a mHMM while for PaGQ_2 from the corresponding mixed parameter model. In both cases, locations are centred and scaled only once, at the beginning of the estimation algorithm. By looking at Tables 3 and 4, it is clear that the former method provides results with a reduced variability than those obtained from the standard approach but with some bias issues. Probably, these are due to flat likelihood surfaces that do not allow to correctly identify the posterior modes and curvatures of the random parameters $\mathbf{b}_i$. On the other hand, PaGQ_2 seems to overcome such problems leading to better results both in terms of bias and standard deviation. This can somehow be explained by the simplified likelihood function that has to be optimized in the internal sub-routine of the algorithm (i.e. the likelihood of a mixed model in place of the likelihood of a mHMM). As a result, global maxima are easier to be detected and standard abscissas may be more appropriately centred and scaled. Based on these findings, we may conclude that PaGQ_2 may be preferred with respect to PaGQ_1. Clearly, the gap between GQ and PaGQ_2 (but also between GQ and PaGQ_1) is bigger for parameters which are directly related to the $\mathbf{b}_i$'s, that is to the longitudinal model parameter (in particular ϕ_1 and ϕ_2) and the variance component estimates.

Figure 1: Response variable distribution for the Poisson mHMM



5.4. Simulation results - the Bernoulli case

Tables 5 and 6 report simulation results for the Bernoulli case. In such a scenario, the optimization routine required to derive parameter estimates and the sub-routine required to compute the posterior modes and curvatures of the random parameter distribution are quite demanding. We have done only some empirical attempts to calculate estimates under the fully adaptive approach that, however, is not feasible in this context due to the high computational times: GQ and PaGQ seem the only viable ways to approximate the intractable integrals that characterize mHMMs. Moreover, as regards the pseudo-adaptive approximations, based on the findings of the Poisson data scenario described in Section 5.3, we have decided to show results only for the PaGQ_2 approach.

As it can be easily noticed, the quality of results obtained under such a scenario is lower than that observed for the Gaussian and the Poisson case. This reflects the higher complexity of the likelihood function to be optimized to derive parameter estimates. Comparing the standard and the pseudo-adaptive approximation, it can be noticed that, when a reduced sample size is available, the latter outperforms the standard approach in most of the cases, especially in terms of standard deviation. As expected, differences may be mainly observed for the longitudinal model parameters and for the variance components, while no relevant differences can be found for the parameters of the hidden Markov chain. Some exceptions are, however, present, mainly for the parameters in the longitudinal data model associated with the

Table 3: Simulation study. Mean parameter estimates for the Poisson mHMM with $n = 250, m = 2$ and $G = 7$

T = 6									
		GQ		aGQ		PaGQ_1		PaGQ_2	
		BIAS	SD	BIAS	SD	BIAS	SD	BIAS	SD
δ_1	0.2	0.085	0.111	0.077	0.098	0.097	0.115	0.086	0.113
δ_2	0.8	-0.085	0.111	-0.077	0.098	-0.097	0.115	-0.086	0.113
q_{11}	0.7	-0.065	0.154	-0.049	0.142	-0.113	0.151	-0.082	0.140
q_{12}	0.3	0.065	0.154	0.049	0.142	0.113	0.151	0.082	0.140
q_{21}	0.1	0.093	0.119	0.080	0.119	0.129	0.128	0.103	0.121
q_{22}	0.9	-0.093	0.119	-0.080	0.119	-0.129	0.128	-0.103	0.121
ϕ_0	-0.8	-0.018	0.154	-0.017	0.121	-0.036	0.139	-0.040	0.128
ϕ_1	-0.6	0.079	0.239	0.016	0.139	0.136	0.189	0.040	0.160
ϕ_2	0.4	0.019	0.391	0.007	0.177	-0.008	0.205	0.005	0.176
β_{11}	0.5	-0.093	0.245	-0.071	0.181	-0.142	0.230	-0.092	0.201
β_{21}	0.7	-0.102	0.250	-0.079	0.211	-0.153	0.256	-0.119	0.241
β_{12}	0.1	0.014	0.163	0.004	0.154	0.045	0.170	0.016	0.154
β_{22}	0.2	0.017	0.179	0.000	0.144	0.041	0.166	0.023	0.160
σ_{11}	1.0	0.177	0.344	0.014	0.216	0.184	0.280	0.040	0.234
σ_{12}	0.5	0.006	0.268	-0.022	0.227	-0.028	0.250	-0.018	0.228
σ_{22}	2.0	0.200	0.577	-0.096	0.448	0.092	0.566	0.046	0.470
T = 10									
		GQ		aGQ		PaGQ_1		PaGQ_2	
		BIAS	SD	BIAS	SD	BIAS	SD	BIAS	SD
δ_1	0.2	0.080	0.102	0.043	0.080	0.066	0.088	0.047	0.080
δ_2	0.8	-0.080	0.102	-0.043	0.080	-0.066	0.088	-0.047	0.080
q_{11}	0.7	-0.014	0.091	-0.014	0.086	-0.057	0.091	-0.042	0.085
q_{12}	0.3	0.014	0.091	0.014	0.086	0.057	0.091	0.042	0.085
q_{21}	0.1	0.062	0.085	0.033	0.066	0.069	0.070	0.048	0.066
q_{22}	0.9	-0.062	0.085	-0.033	0.066	-0.069	0.070	-0.048	0.066
ϕ_0	-0.8	-0.026	0.137	-0.011	0.095	-0.032	0.103	-0.037	0.094
ϕ_1	-0.6	0.104	0.254	0.016	0.114	0.170	0.191	0.038	0.136
ϕ_2	0.4	0.034	0.372	0.008	0.162	-0.014	0.206	0.008	0.165
β_{11}	0.5	-0.064	0.178	-0.028	0.114	-0.052	0.122	-0.044	0.131
β_{21}	0.7	-0.068	0.188	-0.032	0.140	-0.073	0.162	-0.044	0.127
β_{12}	0.1	-0.001	0.132	-0.003	0.091	0.003	0.094	0.001	0.087
β_{22}	0.2	-0.008	0.139	-0.011	0.089	0.004	0.098	-0.004	0.080
σ_{11}	1.0	0.174	0.291	0.011	0.180	0.204	0.284	0.031	0.194
σ_{12}	0.5	0.021	0.231	-0.010	0.178	-0.020	0.202	-0.007	0.180
σ_{22}	2.0	0.235	0.540	-0.156	0.374	0.002	0.456	-0.049	0.378

Table 4: Simulation study. Mean parameter estimates for the Poisson mHMM with $n = 500, m = 2$ and $G = 7$

T = 6									
		GQ		aGQ		$PaGQ_1$		$PaGQ_2$	
		BIAS	SD	BIAS	SD	BIAS	SD	BIAS	SD
δ_1	0.2	0.086	0.099	0.081	0.082	0.104	0.098	0.083	0.086
δ_2	0.8	-0.086	0.099	-0.081	0.082	-0.104	0.098	-0.083	0.086
q_{11}	0.7	-0.044	0.114	-0.020	0.091	-0.070	0.103	-0.054	0.089
q_{12}	0.3	0.044	0.114	0.020	0.091	0.070	0.103	0.054	0.089
q_{21}	0.1	0.079	0.090	0.064	0.075	0.105	0.084	0.078	0.063
q_{22}	0.9	-0.079	0.090	-0.064	0.075	-0.105	0.084	-0.078	0.063
ϕ_0	-0.8	-0.013	0.118	-0.005	0.084	-0.023	0.090	-0.027	0.084
ϕ_1	-0.6	0.061	0.184	0.022	0.098	0.114	0.136	0.036	0.108
ϕ_2	0.4	0.021	0.310	0.004	0.126	-0.010	0.151	0.003	0.137
β_{11}	0.5	-0.087	0.217	-0.063	0.155	-0.106	0.179	-0.074	0.132
β_{21}	0.7	-0.100	0.223	-0.067	0.136	-0.120	0.168	-0.096	0.149
β_{12}	0.1	0.004	0.133	-0.008	0.100	0.013	0.109	0.001	0.088
β_{22}	0.2	0.007	0.135	-0.021	0.096	0.009	0.115	-0.003	0.095
σ_{11}	1	0.150	0.207	0.024	0.147	0.152	0.208	0.037	0.151
σ_{12}	0.5	0.004	0.174	-0.021	0.147	-0.037	0.181	-0.018	0.148
σ_{22}	2	0.195	0.430	-0.127	0.327	0.032	0.412	0.017	0.319
T = 10									
		GQ		aGQ		$PaGQ_1$		$PaGQ_2$	
		BIAS	SD	BIAS	SD	BIAS	SD	BIAS	SD
δ_1	0.2	0.068	0.084	0.041	0.057	0.066	0.072	0.047	0.060
δ_2	0.8	-0.068	0.084	-0.041	0.057	-0.066	0.072	-0.047	0.060
q_{11}	0.7	-0.012	0.069	-0.004	0.055	-0.049	0.061	-0.038	0.060
q_{12}	0.3	0.012	0.069	0.004	0.055	0.049	0.061	0.038	0.060
q_{21}	0.1	0.051	0.071	0.024	0.034	0.063	0.044	0.042	0.035
q_{22}	0.9	-0.051	0.071	-0.024	0.034	-0.063	0.044	-0.042	0.035
ϕ_0	-0.8	-0.020	0.105	-0.001	0.061	-0.027	0.080	-0.038	0.067
ϕ_1	-0.6	0.082	0.220	0.022	0.084	0.180	0.161	0.034	0.093
ϕ_2	0.4	0.023	0.333	0.009	0.116	-0.024	0.169	0.000	0.123
β_{11}	0.5	-0.044	0.134	-0.023	0.062	-0.056	0.088	-0.043	0.090
β_{21}	0.7	-0.050	0.134	-0.026	0.067	-0.061	0.085	-0.041	0.064
β_{12}	0.1	-0.004	0.108	-0.005	0.048	0.008	0.062	0.003	0.057
β_{22}	0.2	-0.014	0.104	-0.016	0.048	-0.004	0.065	-0.006	0.047
σ_{11}	1.0	0.143	0.240	0.001	0.121	0.195	0.202	0.021	0.123
σ_{12}	0.5	0.005	0.166	-0.022	0.129	-0.048	0.155	-0.025	0.126
σ_{22}	2.0	0.247	0.453	-0.152	0.265	0.016	0.345	-0.029	0.264

first hidden state (that progressively becomes less referenced), the covariance between the random parameters b_{i1} and b_{i2} and the variance of b_{i2}. In these cases, estimates obtained through PaGQ$_2$ are more biased but less variable than those obtained through the standard GQ approach. All in all, RMSE is still lower for the former method, but this points out that the adaptive scheme needs more "signal" to work satisfactorily. As regards the variance components estimates $(\sigma_{12}, \sigma_{22})$, with increasing T, the pseudo-adaptive estimates fill almost completely the gap with those obtained through GQ ensuring lower standard deviation across samples. As it will be discussed in Section 5.5, under the GQ approach, relevant (non) convergence issues have been encountered, leading to a reduced number of valid samples. For this reason, to ensure the comparability of results obtained under the two approximation methods, we have decided to enlarge the number of simulated samples generated under the GQ approach in order to get $B = 500$. This can somehow influence the simulation results presented in Table 5: the bias and the standard deviation of parameters under the GQ approach have been calculated for the "lucky samples" only, while, for the PaGQ$_2$ approach, we have taken into account also the "unlucky" ones.
When focusing on the simulation scenarios with $n = 500$, the PaGQ$_2$ method seems to outperform the GQ one, especially with increasing T. More in detail, a reduced bias may be observed for the parameters in the longitudinal data model and, more substantially, for the variance components. As far as the standard deviation of parameter estimates is concerned, simulation results highlight the presence of more concentrated estimates when applying PaGQ$_2$ in place of GQ for almost all the parameter estimates and all the considered simulation scenarios. As before, estimated initial and transition probabilities seem not to be influenced by the method used for the integral approximations; as a result, no significant differences may be observed between GQ and PaGQ$_2$.

5.5. Simulation results - computational complexity and convergence

Figures 2-4 show the distribution of the CPU time required for convergence in the most computationally intensive scenario ($n = 500$, $T = 10$), for the Gaussian, the Poisson and the Bernoulli case, respectively. We did not fix the maximum number of iterations but rather check for convergence by looking at the relative increment of the log-likelihood between consecutive iterations. The convergence threshold was set at $\epsilon = 1e - 6$. All CPU times refer to an Intel I5 architecture (3.3 Ghz).

Table 5: Simulation study. Mean parameter estimates for the Bernoulli mHMM with $n = 250, m = 2$ and $G = 7$

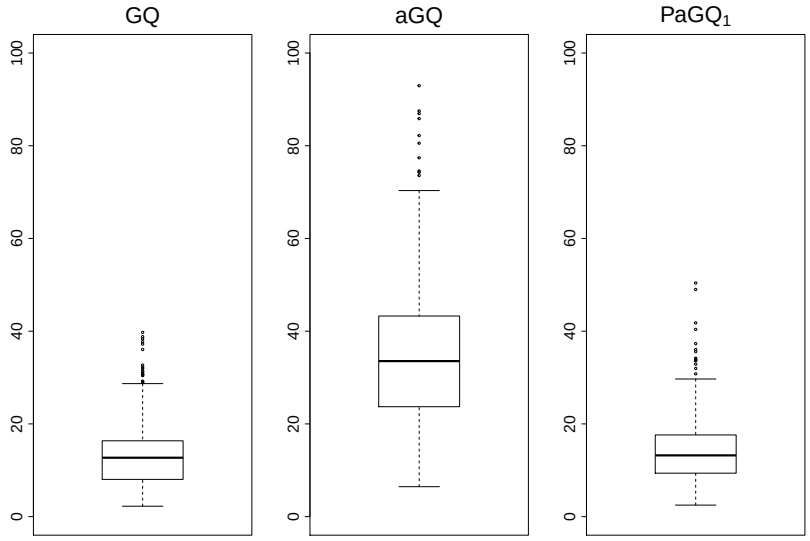
T = 6		GQ		PaGQ_2	
		BIAS	SD	BIAS	SD
δ_1	0.2	0.160	0.204	0.150	0.207
δ_2	0.8	-0.160	0.204	-0.150	0.207
q_{11}	0.7	-0.012	0.171	-0.020	0.183
q_{12}	0.3	0.012	0.171	0.020	0.183
q_{21}	0.1	0.146	0.188	0.161	0.200
q_{22}	0.9	-0.146	0.188	-0.161	0.200
ϕ_0	0.5	-0.011	0.223	0.004	0.203
ϕ_1	-0.7	-0.090	0.358	-0.005	0.333
ϕ_2	-0.6	-0.032	0.210	0.010	0.179
β_{11}	0.5	0.176	0.361	0.190	0.332
β_{21}	1.5	0.531	1.224	0.195	0.707
β_{12}	-1.4	0.096	0.853	0.239	0.708
β_{22}	-0.4	-0.146	1.387	-0.098	0.755
σ_{11}	1	0.665	1.064	0.239	0.997
σ_{12}	0.5	-0.028	0.623	-0.114	0.490
σ_{22}	2	0.137	0.903	-0.256	0.765
T = 10		GQ		PaGQ_2	
		BIAS	SD	BIAS	SD
δ_1	0.2	0.096	0.165	0.099	0.190
δ_2	0.8	-0.096	0.165	-0.099	0.190
q_{11}	0.7	0.000	0.126	0.006	0.136
q_{12}	0.3	0.000	0.126	-0.006	0.136
q_{21}	0.1	0.067	0.123	0.066	0.129
q_{22}	0.9	-0.067	0.123	-0.066	0.129
ϕ_0	0.5	0.003	0.177	0.012	0.170
ϕ_1	-0.7	-0.027	0.287	0.019	0.281
ϕ_2	-0.6	-0.006	0.180	0.009	0.159
β_{11}	0.5	0.088	0.265	0.089	0.263
β_{21}	1.5	0.258	0.556	0.129	0.357
β_{12}	-1.4	0.004	0.634	0.084	0.612
β_{22}	-0.4	0.015	0.310	-0.010	0.320
σ_{11}	1	0.248	0.647	0.056	0.683
σ_{12}	0.5	-0.004	0.442	-0.071	0.372
σ_{22}	2	0.104	0.659	-0.129	0.628

Table 6: Simulation study. Mean parameter estimates for the Bernoulli mHMM with $n = 500, m = 2$ and $G = 7$

T = 10		GQ		PaGQ_2	
		BIAS	SD	BIAS	SD
δ_1	0.2	0.118	0.149	0.114	0.155
δ_2	0.8	-0.118	0.149	-0.114	0.155
q_{11}	0.7	-0.012	0.130	-0.012	0.131
q_{12}	0.3	0.012	0.130	0.012	0.131
q_{21}	0.1	0.099	0.141	0.109	0.157
q_{22}	0.9	-0.099	0.141	-0.109	0.157
ϕ_0	0.5	-0.004	0.156	-0.004	0.152
ϕ_1	-0.7	-0.028	0.236	0.006	0.231
ϕ_2	-0.6	-0.008	0.140	0.014	0.129
β_{11}	0.5	0.140	0.248	0.142	0.260
β_{21}	1.5	0.260	0.499	0.133	0.426
β_{12}	-1.4	0.102	0.549	0.194	0.524
β_{22}	-0.4	0.032	0.309	0.011	0.264
σ_{11}	1	0.296	0.688	0.111	0.684
σ_{12}	0.5	-0.042	0.399	-0.107	0.352
σ_{22}	2	0.107	0.604	-0.144	0.596

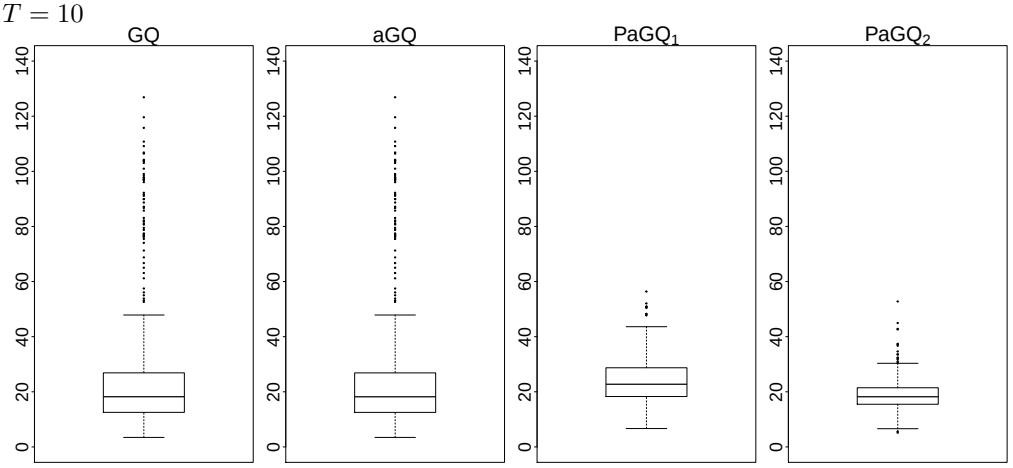
T = 10		GQ		PaGQ_2	
		BIAS	SD	BIAS	SD
δ_1	0.2	0.085	0.120	0.076	0.122
δ_2	0.8	-0.085	0.120	-0.076	0.122
q_{11}	0.7	0.011	0.089	0.016	0.087
q_{12}	0.3	-0.011	0.089	-0.016	0.087
q_{21}	0.1	0.047	0.073	0.039	0.070
q_{22}	0.9	-0.047	0.073	-0.039	0.070
ϕ_0	0.5	0.006	0.123	0.006	0.120
ϕ_1	-0.7	-0.022	0.181	-0.008	0.178
ϕ_2	-0.6	-0.006	0.122	0.005	0.109
β_{11}	0.5	0.089	0.179	0.079	0.176
β_{21}	1.5	0.173	0.248	0.111	0.208
β_{12}	-1.4	0.074	0.380	0.102	0.342
β_{22}	-0.4	0.033	0.309	0.025	0.258
σ_{11}	1	0.141	0.461	0.013	0.456
σ_{12}	0.5	-0.017	0.301	-0.052	0.276
σ_{22}	2	0.065	0.425	-0.073	0.365

Figure 2: Computational time (in minutes) for the Gaussian mHMM, with $n = 500$ and $T = 10$



As it can be observed, in Gaussian and Poisson data scenarios, the fully adaptive quadrature scheme leads to a substantial increase in computational times. In the former case, 12.84, 33.71 and 13.64 minutes (on average) are required to reach the convergence by the standard, the adaptive and the pseudo-adaptive quadrature rule (PaGQ_1), respectively. When fitting the model for conditional Poisson responses, the algorithms reach the convergence in 25.85, 48.86, 23.89 and 18.74 minutes (on average) when GQ, aGQ, PaGQ_1 and PaGQ_2 are applied, respectively. In this simulation scenario, the non linear nature of the longitudinal score function consistently increases times to convergence. While in the Gaussian case the standard and the fully adaptive GQ routines did not pose any problem, in the Poisson case we faced some convergence issues represented by the long right tail of the corresponding distributions of computational times.
In this latter scenario, as far as the pseudo-adaptive routines are concerned, see Figure 3, several comments are possible. The median CPU time required to converge is, somehow, comparable to those of the standard GQ approach. However, it is worth noticing that, in such a framework, PaGQ approaches do not have the long right tails that, instead, can be observed both for the standard and the fully adaptive scheme. The reasons behind such results are

Figure 3: Computational time (in minutes) for the Poisson mHMM, with $n = 500$ and $T = 10$



twofold. On one hand, the GQ approach faces some difficulties in providing reliable estimates for the random parameter covariance matrix and, thus, an increase in the time to convergence occurs. It is interesting to highlight that, because of these difficulties, some of the simulated samples have been discarded from the analysis since the GQ algorithm did not reach convergence, even if this happened in a very limited number (2.4%) of cases. On the other hand, PaGQ schemes avoid the internal optimization routine invoked at each step of the EM algorithm when the aGQ approximation is used. This optimization can be quite complex due to the shape of the likelihood function surface. In particular, under the aGQ approach, we have observed a questionable performance for $\widehat{\mathbf{D}}$ at the initial iterations of the EM algorithm. For this reason, a higher number of iterations is required to move the estimates in the right direction. As regards the comparison between PaGQ_1 and PaGQ_2, it is interesting to notice that using posterior modes and curvatures from a mixed model clearly allows to better initialize the algorithm. This seems to converge in a reduced number of iterations as it is clear from the more concentrated distribution in the fourth panel of Figure 3, when compared to the first and the third ones.

As far as the Bernoulli data scenario is concerned, see Figure 4, it can be noticed that the two routines we have compared (GQ and PaGQ_2) do not present relevant differences in terms of computational load. In both cases, the time to convergence is consistently higher than those observed for the Pois-

Figure 4: Computational time (in minutes) for the Bernoulli mHMM, with $n = 500$ and $T = 10$

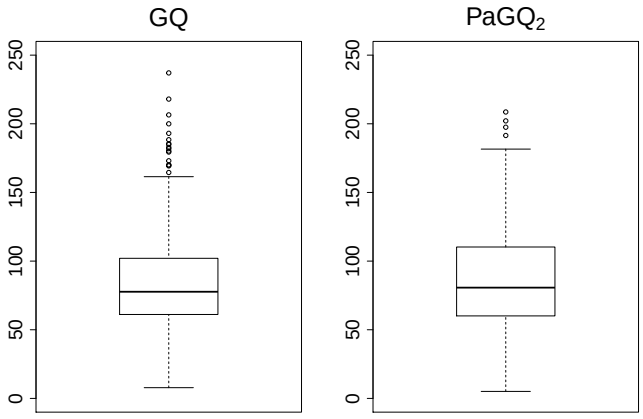


Table 7: Proportion of samples discarded from the analysis due to convergence issues

	GQ	PaGQ_2
n = 250, T = 6	28.4	4.86
n = 250, T = 10	16.92	3.29
n = 500, T = 6	14.80	1.20
n = 500, T = 10	6.80	0.20

son and the Gaussian data scenarios. In fact, on average, 87.77 and 88.22 minutes are required to compute parameter estimates when applying GQ and PaGQ_2, respectively: as we move far from (conditional) Gaussianity, the parameter estimation for mHMMs progressively becomes quite a demanding task. Besides these considerations, the pseudo-adaptive approach still turns out to be more convenient. In Table 7, we show the proportion of samples that have been discarded from the analysis due to lack of convergence or to numerical errors caused by improper values of the variance components. As it is clear, the pseudo-adaptive approach is significantly more under control than the GQ one; this latter approach experiments relevant convergence issues, especially when small sample sizes are considered. Similar to the Poisson case, these are often due to variance component values that go towards zero. This finding can somehow justify the similarities between GQ and PaGQ_2 in terms of computational load in this simulation scenario. This

is in contrast with what we have found for the Poisson data case. Indeed, for the Bernoulli case, only the "lucky" samples have been taken into consideration for the GQ approximation; therefore, the algorithm seems to reach the convergence in a reduced time when this latter approximation is employed. On the other hand, results observed for PaGQ_2 may be influenced by the "unlucky" samples that, instead, require more iterations to converge.

6. Concluding remarks

In this paper, approximations based on different Gaussian quadrature schemes have been compared for ML estimation in mHMMs. The basic assumption underlying these models is that unobserved heterogeneity may lead to dependence between measures recorded on the same individual. This dependence may be represented through zero-mean Gaussian random parameters, as in mixed models, or may be represented through a time-varying latent variable with a Markov-type structure, as in HMMs. When both are present, mHMMs arise. In such a framework, parameter estimation requires the calculation of multiple integrals which, generally, have to be numerically evaluated.

In the present work, standard Gaussian quadrature has been compared with adaptive and pseudo-adaptive quadrature schemes. The latter approaches aim at improving the quality of the approximation by centring and scaling standard locations via posterior modes and curvatures of the random parameter distribution. In the adaptive scheme, this transformation is performed at each step of the algorithm, while in the pseudo-adaptive one, it is performed only once, at the first iteration. To our knowledge, this is the first effort to use adaptive schemes in mHMMs and to compare the quadrature schemes in a large scale simulation study.

Based on the simulation results, we can conclude that the adaptive Gaussian quadrature scheme consistently reduces the approximation error of the standard Gaussian quadrature approximation. This improvement is more evident as we move far from (conditionally) Gaussian responses; in these cases, Gauss-Hermite locations do not allow to properly identify where the main mass of the integrand is located. On the other hand, the computational load required to derive parameter estimates consistently grows, making such an approach quite demanding. Simulation results show that the pseudo-adaptive schemes represent an interesting alternative to both the standard and the fully adaptive approximation when a sufficient number of repeated

measurements per unit is available (here $T \geq 6$). In this case, the quality of the parameter estimates obtained by centring and scaling standard quadrature locations only at the first iteration of the EM algorithm results in a clear improvement over GQ while having a similar computational effort.
In our development, we have taken into consideration the presence of varying number of time measurements for each sample unit. A common problem to deal with is the presence of a potentially informative drop-out, i.e. of a missing data generation process that may be influenced by both observed and unobserved longitudinal responses. In such a situation, the missing data mechanism has to be taken into account to obtain consistent estimates. A reference is the shared parameter model proposed by Bartolucci and Farcomeni (2015) where a mHMM based on time-constant and time-varying intercepts is applied to multivariate responses in the presence of (discrete time) drop-out. A further proposal in the HMM literature has been introduced by Maruotti (2015) within the class of pattern mixture models.

Appendix A. Calculation of posterior modes and curvatures

As it has been pointed out in the previous sections, estimation of model parameters requires the calculation of multiple integrals which, often, have not a closed form and require the use of numerical approximations to be evaluated. Gaussian quadrature represents a possible solution. Standard Gauss-Hermite rule approximates integrals through a weighted sum over a pre-specified set of abscissas. However, as emphasised before, when the integrand function is not centred at zero, the approximation may fail, even for a large number of quadrature points (see e.g. Rizopoulos 2012). To solve the problem, an adaptive quadrature scheme can be used instead: a linear transformation of the standard abscissas is applied to improve the quality of the approximation. More precisely, the mode and the curvature of the (individual-specific) log integrand are calculated at each step of the EM algorithm and used to scale and shift standard Gauss-Hermite abscissas placing the peak of the integrand function around zero.
In this section, we examine in detail the computation of these quantities. To simplify the notation, all the subscripts in the probability density functions are suppressed. Starting from the observed data likelihood

$$\int \sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i) \mathrm{d}\mathbf{b}_i,$$

we may notice that the integrand function is proportional to the posterior distribution of the random vector $\mathbf{b}_i$; therefore, we may approximate the integrand with a quantity having the same (log) derivatives up to second order. To this purpose, posterior modes and curvatures of $\mathbf{b}_i$ have to be computed. In the following, the first order derivative of the integrand, on the logarithmic scale, is derived.

$$\begin{aligned}
&\frac{\partial \log\left[\sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)\right]}{\partial \mathbf{b}_i} \\
&= \frac{1}{\sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)} \sum_{\mathbf{s}_i} \frac{\partial f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i} \\
&= \sum_{\mathbf{s}_i} \frac{f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)}{\sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)} \frac{\partial \log f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i} \\
&= \sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \frac{\partial \log f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i}. \qquad \text{(A.1)}
\end{aligned}$$

From expression (A.1), it is clear that the mode of the random coefficient vector is obtained by finding the zeros of a weighted mean: the complete score function of each unit is weighted by the distribution of the latent markovian variables, given the observed data and the random coefficients.
Based on the modelling assumptions introduced in section 2, the complete data score is given by

$$\begin{aligned}
&\frac{\partial \log f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i} \\
&= \frac{\partial}{\partial \mathbf{b}_i} \log \left[\prod_{t=1}^{T_i} f(y_{it} \mid s_{it}, \mathbf{b}_i) q(s_{it} \mid s_{it-1}) f(\mathbf{b}_i) \right]
\end{aligned}$$

where, for the sake of simplicity, we assume $q(s_{i1} \mid s_{i0}) = \delta(s_{i1})$. Omitting all the components that do not depend on the random vector $\mathbf{b}_i$ and indicating with $\mu_{s_{it}}$ and $v_{s_{it}}$ the response mean value and the variance function for a

generic subject i being, at time t, in state s_{it}, we get:

$$\begin{aligned}
&\frac{\partial \log f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i} \\
&= \sum_{t=1}^{T_i} \frac{\partial}{\partial \mathbf{b}_i} \log f(y_{it} \mid s_{it}, \mathbf{b}_i) + \frac{\partial}{\partial \mathbf{b}_i} \log f(\mathbf{b}_i) \\
&= \sum_{t=1}^{T_i} \mathbf{z}_{it} \left[\frac{\partial \mu_{s_{it}}}{\partial \eta_{s_{it}}} \right] v_{s_{it}}^{-1} \left[y_{it} - \mu_{s_{it}} \right] - \mathbf{D}^{-1} \mathbf{b}_i, \qquad \text{(A.2)}
\end{aligned}$$

Inserting the above results in (A.1) and indicating with $e_{s_{it}}$ the model residuals $[y_{it} - \mu_{s_{it}}]$, the following expression holds

$$\begin{aligned}
&\frac{\partial \log \left[\sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i) \right]}{\partial \mathbf{b}_i} \\
&= \sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \left\{ \sum_{t=1}^{T_i} \mathbf{z}_{it} \left[\frac{\partial \mu_{s_{it}}}{\partial \eta_{s_{it}}} \right] v_{s_{it}}^{-1} e_{s_{it}} - \mathbf{D}^{-1} \mathbf{b}_i \right\}. \qquad \text{(A.3)}
\end{aligned}$$

If a canonical link is employed, the previous equation simplifies to:

$$\sum_{\mathbf{s}_i} f(\mathbf{s}_i, \mid \mathbf{y}_i, \mathbf{b}_i) \left\{ \sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1} \mathbf{b}_i \right\} \qquad \text{(A.4)}$$

since, in this case, we have $\frac{\partial \mu_{s_{it}}}{\partial \eta_{s_{it}}} = \frac{\partial \mu_{s_{it}}}{\partial \theta_{s_{it}}} = v_{s_{it}}$.
Considering (A.4) as the score function, we may derive the negative Hessian as follows:

$$\begin{aligned}
\mathbf{H}_i &= -\frac{\partial \log \left[\sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i) \right]}{\partial \mathbf{b}_i \mathbf{b}_i^{\mathsf{T}}} \\
&= -\sum_{\mathbf{s}_i} \left[\sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1} \mathbf{b}_i \right] \left[\frac{\partial f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i^{\mathsf{T}}} \right] + \\
&\quad - \sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \frac{\partial \left\{ \sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1} \mathbf{b}_i \right\}}{\partial \mathbf{b}_i^{\mathsf{T}}},
\end{aligned}$$

which, as it can be easily observed, resembles the i-th unit contribution to the observed information matrix in the Louis (1982) formula. By straightforward

calculation, we obtain:

$$\begin{aligned}\mathbf{H}_i = &- \sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \\ &\times \left[\sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1}\mathbf{b}_i\right] \left[\frac{\partial \log\left[f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i)\right]}{\partial \mathbf{b}_i^{\mathsf{T}}}\right] \\ &+ \sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \left[\sum_{t=1}^{T_i} \mathbf{z}_{it} v_{s_{it}} \mathbf{z}_{it}^{\mathsf{T}} + \mathbf{D}^{-1}\right], \end{aligned} \tag{A.5}$$

where the second term is the posterior mean (with respect to $f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i)$) of the complete-data information on $\mathbf{b}_i$. By doing a little algebra, the first term in expression (A.5) can be computed as follows:

$$\begin{aligned}\mathbf{A}_i = &- \sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \\ &\times \left[\sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1}\mathbf{b}_i\right] \left[\frac{\partial \log\left[f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i)\right]}{\partial \mathbf{b}_i^{\mathsf{T}}}\right] \\ = &- \left\{\sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \left[\sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1}\mathbf{b}_i\right]\right. \\ &\times \left.\left[\sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1}\mathbf{b}_i\right]^{\mathsf{T}}\right\} + \left\{\sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i)\right. \\ &\times \left.\left[\sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1}\mathbf{b}_i\right] \left[\frac{\partial \log \sum_{\mathbf{s}_i} f(\mathbf{y}_i, \mathbf{s}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i^{\mathsf{T}}}\right]\right\}. \end{aligned}$$

The last term corresponds to expressions (A.3) and (A.4), i.e. the first derivative of the log joint distribution of $\mathbf{b}_i$ and $\mathbf{y}_i$. When calculated at $\hat{\mathbf{b}}_i$, this term is null and the individual-specific negative Hessian matrix $\mathbf{H}_i$ is given

by the following expression:

$$\mathbf{H}_i = -\sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \left\{ \left[\sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1}\mathbf{b}_i \right] \times \left[\sum_{t=1}^{T_i} \mathbf{z}_{it} e_{s_{it}} - \mathbf{D}^{-1}\mathbf{b}_i \right]^{\mathsf{T}} \right\} + \sum_{\mathbf{s}_i} f(\mathbf{s}_i \mid \mathbf{y}_i, \mathbf{b}_i) \left[\sum_{t=1}^{T_i} \mathbf{z}_{it} v_{s_{it}} \mathbf{z}_{it}^{\mathsf{T}} + \mathbf{D}^{-1} \right]. \quad \text{(A.6)}$$

In the adaptive quadrature approach, posterior modes and curvatures derived in this appendix are used to modify standard Gaussian quadrature points, at each step of the EM algorithm. As outlined in section 4, when a pseudo-adaptive rule is preferred, a linear transformation of standard abscissas is applied by shifting and scaling the standard locations by using posterior modes and curvatures computed only at the beginning of the optimization routine, according to the expressions given in this appendix.

References

Aitkin, M., 1999. A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55, 117–128.

Altman, R. J., 2007. Mixed hidden Markov models: an extension of the hidden markov model to the longitudinal data setting. Journal of the American Statistical Association 102 (477), 201–210.

Bartolucci, F., Farcomeni, A., 2015. A discrete time event-history approach to informative drop-out in mixed latent markov models with covariates. Biometrics, in press.

Bartolucci, F., Farcomeni, A., Pennoni, F., 2012. Latent Markov Models for Longitudinal Data. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. Taylor & Francis.

Baum, L. E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics 41, 164–171.

Böhning, D., 1982. Convergence of simar's algorithm for finding the maximum likelihood estimate of a compound poisson process. The Annals of Statistics 10, 1006–1008.

Cagnone, S., Monari, P., 2013. Latent variable models for ordinal data by using the adaptive quadrature approximation. Computational Statistics 28 (2), 597–619.

Cappé, O., Moulines, E., Ryden, T., 2005. Inference in hidden Markov models. Springer Series in Statistics. Springer.

Chaubert-Pereira, F., Gudon, Y., Lavergne, C., Trottier, C., 2010. Markov and semi-markov switching linear mixed models used to identify forest tree growth components. Biometrics 66, 753–762.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B 39, 1–38.

Diggle, P. J., Heagerty, P. J., Liang, K. Y., Zeger, S. L., 2002. Analysis of longitudinal data, 2nd Edition. Vol. 25 of Oxford Statistical Science Series. Oxford University Press.

Hartzel, J., Agresti, A., Caffo, B., 2001. Multinomial logit random effects models. Statistical Modelling 1, 81–102.

Heckman, J. J., 1981. The incidental parameters problem and the problem of initial conditions in estimating discrete time-discrete data stochastic processes and some Monte Carlo evidence. In: Manski, C., McFadden, D. (Eds.), Structural analysis of discrete data. MIT Press.

Lagona, F., Jdanov, D., Shkolnikova, M., 2014. Latent time-varying factors in longitudinal analysis: a linear mixed hidden markov model for heart rates. Statistics in Medicine 33, 4116–4134.

Laird, N., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association 73, 805–811.

Laird, N. M., Ware, J. H., 1982. Random-effects models for longitudinal data. Biometrics 38, 963–974.

Langrock, R., Hopcraft, G., Blackwell, P., Goodall, V., King, R., Niu, M., Patterson, T., Pedersen, M., Skarin, A., Schick, R. S., 2014. Modelling group dynamic animal movement. Methods in Ecology and Evolution 5, 190–199.

Lindsay, B. G., 1983a. The geometry of mixture likelihoods: a general theory. Ann. Statist. 11, 86–94.

Lindsay, B. G., 1983b. The geometry of mixture likelihoods, part ii: the exponential family. The Annals of Statistics 11, 783–792.

Liu, Q., Pierce, D. A., 1994. A note on Gauss-Hermite quadrature. Biometrika 81, 624–629.

Louis, T. A., 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 44, 226–233.

Maruotti, A., 2011. Mixed hidden markov models for longitudinal data: An overview. International Statistical Review 79, 1751–5823.

Maruotti, A., 2015. Handling non-ignorable dropouts in longitudinal data: a conditional model based on a latent markov heterogeneity structure. TEST 24, 84–109.

Maruotti, A., Rydén, T., 2009. A semiparametric approach to hidden markov models under longitudinal observations. Statistics and Computing 19 (4), 381–393.

Neuhaus, J. M., McCulloch, C. E., Boylan, R., 2013. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercept and slopes. Statistics in Medicine 32, 2419–2429.

Patterson, H. D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545–554.

Pinheiro, J. C., Bates, D. M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics 4, 12–35.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. Stata Journal 2, 1–21.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2005. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. Journal of Econometrics 128, 301 – 323.

Rijmen, F., Vansteelandt, K., Boeck, P., 2008. Latent class models for diary method data: Parameter estimation by local computations. Psychometrika 73, 167–182.

Rizopoulos, D., 2012. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. Computational Statistics and Data Analysis 56, 491–501.

Zucchini, W., MacDonald, I. L., 2009. Hidden Markov models for time series. Vol. 110 of Monographs on Statistics and Applied Probability. CRC Press.