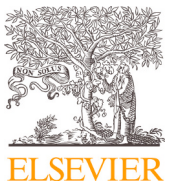
Computational Statistics and Data Analysis 222 (2026) 108402



Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A penalized maximum likelihood approach to deal with latent state separation in hidden Markov models with covariates and lagged responses

Luca Brusa [a,*], Fulvia Pennoni [a], Francesco Bartolucci [b], Romina Peruilh Bagolini [b]

[a] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Milan, 20126, Italy
[b] Department of Economics, University of Perugia, Via Alessandro Pascoli 20, Perugia, 06123, Italy

ARTICLE INFO

Keywords:
Binary longitudinal data
Discrete latent variables
Early-warning system
Expectation-maximization algorithm
Hypotension data

ABSTRACT

A penalized maximum likelihood estimation approach is proposed for discrete-time hidden Markov models in which the manifest distribution depends on covariates and the lagged response. The proposed method addresses the issue of latent state separation, typically arising with binary responses or categorical response variables with a limited number of categories, which leads to extremely large estimates of the support points of the latent variable distribution. We also propose a cross-validation approach for jointly selecting the number of latent states of the model and the strength of the likelihood penalization. The approach is validated through a deep simulation study aimed at comparing parameter estimation accuracy and computational efficiency across different estimation procedures. Finally, we illustrate the proposed approach by analyzing longitudinal data collected during spinal anesthesia, including covariates, with the aim of monitoring the occurrence of hypotension in certain patients.

1. Introduction

In longitudinal studies, the focus typically lies in the evolution over time of a phenomenon of interest assessed through occasion-specific response variables. Moreover, it is often crucial to evaluate how covariates affect the response variables and how this effect changes over time. Relevant issues include identifying the factors that affect the response variables as well as capturing unobserved heterogeneity and its evolution across time. Several models are available for these types of analysis, including hidden Markov (HM) models (Bartolucci et al., 2013; Mor et al., 2021; Visser and Speekenbrink, 2022), also named latent Markov models. These, in particular, are suitable for dynamic clustering of sample units, as they assume, for every unit, the existence of a latent process that affects the distribution of the response variables. Under the assumption of local independence, each process follows a first-order Markov chain with a finite number of states. HM models can also be formulated by including individual covariates either in latent (sub)model, so that they affect the initial and transition probabilities of the latent Markov chain through a suitable parameterization (Bartolucci et al., 2014; Maruotti and Punzo, 2017), or in the measurement (sub)model, so that, with binary or categorical responses, they directly affect the conditional response probabilities. In the first case, the covariates influence the latent state dynamics, thereby shaping the transition process to account for the possible presence of measurement errors (Bartolucci et al., 2013). In the second case, the latent processes are used to capture unobserved heterogeneity in a dynamic fashion. We are referring, in particular, to the

* Corresponding author.
E-mail addresses: luca.brusa@unimib.it (L. Brusa), fulvia.pennoni@unimib.it (F. Pennoni), francesco.bartolucci@unipg.it (F. Bartolucci), romina.peruilh@dottorandi.unipg.it (R. Peruilh Bagolini).

https://doi.org/10.1016/j.csda.2026.108402
Received 3 July 2025; Received in revised form 30 April 2026; Accepted 30 April 2026
Available online 1 May 2026
0167-9473/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

heterogeneity in individual responses that cannot be explained by the observed covariates or, equivalently, to the residual variability after having accounted for all measurable individual characteristics. This is a crucial aspect in the analysis of longitudinal data, as failing to account for it may lead to biased or incomplete conclusions. In many applications, it may be also of interest to account for conditional serial dependence between responses provided by the same individual given the latent process. For this purpose, it is possible to include the lagged response among the covariates, so as to relax the assumption of conditional independence when it is considered too restrictive. This happens in certain applications involving, for instance, capture-recapture data (Bartolucci and Pennoni, 2007).

Considering categorical response variables, a convenient parameterization of the conditional distribution of the responses in a HM model may be based on generalized logits. This specification includes the support points of the latent distribution, namely, parameters measuring the effect of each hidden state on the probabilities of the response categories at every time occasion, as well as regression parameters for the covariates and the lagged response. In this way, the hidden states are easily interpretable in terms of tendency or propensity toward specific behaviors, and the overall model can be viewed as a generalization of the dynamic logit model (Hsiao, 2003) with time-varying and discrete random effects.

Maximum likelihood estimation of the models at issue is carried out relatively straightforwardly using the expectation-maximization (EM) algorithm (Dempster et al., 1977). At each maximization step (M-step), the optimization involving the covariate parameters is performed using the Newton-Raphson (NR) iterative algorithm, initialized at the estimates obtained from the previous expectation step (E-step) of the EM algorithm. To assess the precision of the maximum likelihood estimates, asymptotic standard errors are obtained in the usual way through the observed information matrix, which may be computed numerically (Bartolucci and Farcomeni, 2009). The formulation based on discrete latent variables allows for simpler inferential procedures than alternative latent specifications in which the random effects follow a continuous distribution (see, e.g., Cagnone et al., 2009). However, the estimates of the support point parameters associated with the hidden states may sometimes take extreme values, owing to form a separation problem. Separation problems are well known and have been extensively studied for many statistical models for categorical response variables, such as logistic regression; see, among others, Mansournia et al. (2017) and Clark et al. (2023). In these contexts, a separation problem typically arises when covariates perfectly predict the response variable, potentially leading to infinite estimates for some regression coefficients. In such settings, the causes of separation are well established and generally include small sample sizes, rare outcomes, and highly correlated covariates (Mansournia et al., 2017).

In the context of HM models, separation primarily concerns the support points of the latent states, leading to extremely large estimates in absolute value. As a consequence, the estimated covariate effects may become negligible. To address this issue, in the present paper we propose a penalized likelihood method based on the squared deviations of the support points from their average, with the aim of preventing overly large support point estimates and ensuring reliable statistical inference. In the general framework of nonparametric probability density estimation, the use of roughness penalties in likelihood-based estimation was introduced by Good and Gaskins (1971), who proposed maximizing a log-likelihood penalized by a roughness functional, and was later further developed by Silverman (1982) within the maximum penalized likelihood framework.

Penalized likelihood methods are not new in the HM literature. They have been used for several purposes, including bias reduction (Farcomeni, 2017), selection of the number of latent states (Hung et al., 2013; Lin and Song, 2022), variable selection in the presence of many covariates (Ötting and Groll, 2022), smoothing of nonparametric emission distributions (Langrock et al., 2015), and regularization of unbounded likelihoods in Gaussian HM models (Alexandrovich, 2016). However, to the best of our knowledge, this is the first penalized approach for HM models aimed at regularizing the dispersion of the state-specific support points. The proposed penalty directly targets latent-state separation, while preserving the interpretation of the support points as ordered latent propensities and preventing extreme estimates. From a computational perspective, we show that the resulting penalized maximization can be embedded in the EM algorithm through a simple modification of the M-step, leaving the E-step unchanged. Additionally, we propose a procedure, inspired by Smyth (2000), for cross-validating the likelihood in model-based probabilistic clustering contexts. Specifically, we use cross-validation to jointly select the number of states of the hidden chain and the strength of the penalty (Scott et al., 1978).

To evaluate the performance of the proposed estimator, we conduct an extensive simulation study. We compare penalized and standard maximum likelihood estimation in terms of accuracy and computational time. We also report the results of the cross-validation approach used to select the optimal value of the penalization parameter. Moreover, we illustrate the proposed method through an application to the occurrence of hypotension, measured every five minutes during spinal anesthesia. The identification of patient risk group profiles during the intervention, as well as the assessment of the covariate effects, such as the patient's position during the surgery, remains an open area of research, as discussed by Radwan et al. (2024). We compare our results with those reported by Aktas et al. (2014), who analyzed the same data using a different approach.

The remainder of the paper is organized as follows. In Section 2, we recall the model formulation and maximum likelihood estimation. In Section 3, we present the proposed penalization method and the estimation procedure. In Section 4, we show the results of the simulation study and, in Section 5, we illustrate the results of the application. Finally, in Section 6 we provide some concluding remarks. The estimation procedures are implemented by extending the functions from the `LMest` package (Pennoni et al., 2025), developed for the R software (Team, 2025); the routines are available from https://github.com/LB1304/PenHMM.

2. Model formulation

Let $\boldsymbol{Y}_i = (Y_i^{(1)}, \dots, Y_i^{(T)})'$ and $\boldsymbol{X}_i = (\boldsymbol{X}_i^{(1)}, \dots, \boldsymbol{X}_i^{(T)})'$, where $Y_i^{(t)}$ denotes the univariate binary response variable and $\boldsymbol{X}_i^{(t)}$ the vector of individual covariates for subject i at time t, for $i = 1, \dots, n$ and $t = 1, \dots, T$. Realizations of $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ are denoted by $\boldsymbol{y} = (y^{(1)}, \dots, y^{(T)})'$ and $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(T)})'$, respectively, and a similar convention will be used for other random variables and vectors.

In the following, we illustrate the HM formulation to model the conditional distribution of the response variables given the covariates.

2.1. Hidden Markov model assumptions

For every unit i, let $\boldsymbol{U}_i = (U_i^{(1)}, \dots, U_i^{(T)})'$ be a latent process assumed to follow a first-order Markov chain with a finite number of hidden states $\{1, \dots, k\}$. The HM model is characterized by two components: (i) the measurement (sub)model, whose probability mass (or density) function is denoted as $p_{\boldsymbol{Y}_i|\boldsymbol{U}_i,\boldsymbol{X}_i}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{x})$ and represents the distribution of the response vector $\boldsymbol{Y}_i$ given the latent process $\boldsymbol{U}_i$ and the observable covariates $\boldsymbol{X}_i$, and (ii) the latent (sub)model, whose probability mass function is denoted as $p_{\boldsymbol{U}_i}(\boldsymbol{u})$ and represents the distribution of the latent process. In this formulation, the covariates are assumed to affect solely the measurement (sub)model. In the following, we also refer to the manifest distribution of the response variable given the covariates as

$$p_{\boldsymbol{Y}_i|\boldsymbol{X}_i}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{u}} p_{\boldsymbol{Y}_i|\boldsymbol{U}_i,\boldsymbol{X}_i}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{x}) p_{\boldsymbol{U}_i}(\boldsymbol{u}).$$

To efficiently compute this distribution the Baum-Welch forward recursion is adopted (Baum and Petrie, 1966; Welch, 2003).

In the formulation here adopted, parameters related to the latent sub-model do not depend on covariates and they are: (i) the initial probability of each state, denoted as $\pi_u = p(U_i^{(1)} = u)$, $u = 1, \dots, k$, satisfying $\pi_u \geq 0$ and $\sum_{u=1}^{k} \pi_u = 1$, and (ii) the transition probabilities among hidden states, denoted as $\pi_{u|\bar{u}} = p(U_i^{(t)} = u | U_i^{(t-1)} = \bar{u})$, $\bar{u}, u = 1, \dots, k$, $t = 1, \dots, T$, satisfying $\pi_{u|\bar{u}} \geq 0$ and $\sum_{u=1}^{k} \pi_{u|\bar{u}} = 1$ for all $\bar{u}$. To ensure model identifiability, additional constraints must be imposed, as detailed below. Hidden states correspond to different subpopulations and are associated with different levels of the effect of unobserved covariates on the response variable. In this way, unobserved heterogeneity is accounted for dynamically, as each individual may move between hidden states over time.

According to the local independence assumption, the response variables are assumed to be conditionally independent given $\boldsymbol{U}_i$. This assumption can be relaxed by including the lagged response variable $Y_i^{(t-1)}$ in each covariate vector $\boldsymbol{X}_i^{(t)}$, thereby allowing for serial dependence between observed responses over time even conditionally on the underlying latent process. Suitable parameterizations can be adopted to model the dependence of the response variable on these covariates (Bartolucci et al., 2013). Here, we consider a parameterization in the form of a logit model (McCullagh and Nelder, 1989; Agresti, 2002) specified as follows:

$$\log \frac{p(Y_i^{(t)} = 1 | U_i^{(t)} = u, \boldsymbol{X}_i^{(t)} = \boldsymbol{x})}{p(Y_i^{(t)} = 0 | U_i^{(t)} = u, \boldsymbol{X}_i^{(t)} = \boldsymbol{x})} = \log \frac{\phi_{1|u\boldsymbol{x}}^{(t)}}{\phi_{0|u\boldsymbol{x}}^{(t)}} = \log \frac{\phi_{1|u\boldsymbol{x}}^{(t)}}{1 - \phi_{1|u\boldsymbol{x}}^{(t)}} = \alpha_u + \boldsymbol{x}'\boldsymbol{\beta}, \tag{1}$$

where $\phi_{1|u\boldsymbol{x}}^{(t)}$ and $\phi_{0|u\boldsymbol{x}}^{(t)}$ denote the conditional response probabilities given the hidden state and the covariate configuration $\boldsymbol{x}$, $u = 1, \dots, k$ and $t = 1, \dots, T$. In this expression, $\boldsymbol{\beta}$ is the vector of regression parameters related to the covariates and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)$ is the vector of state-specific support points. Moreover, α_u measures the tendency toward the success event for subjects in state u. To ensure model identifiability, we impose the constraint $\alpha_1 < \alpha_2 < \dots < \alpha_k$. This ordering resolves the label switching problem without restricting likelihood maximization, since the states can be relabeled without loss of generality. A key feature of the model, under the previous assumptions, is the inclusion of the time-varying direct effects of the covariate on the response variable. Overall, the model can be viewed as a discrete version of a random-effects logit model (McCulloch et al., 2008), extending the dynamic logit model proposed in Hsiao (2003).

2.2. Maximum likelihood estimation

On the basis of a sample of n observations, maximum likelihood estimation of the model parameters, collected in the vector $\boldsymbol{\theta}$, is carried out using the EM algorithm (Dempster et al., 1977). This algorithm maximizes the observed-data log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{Y}_i|\boldsymbol{X}_i}(\boldsymbol{y}|\boldsymbol{x}),$$

relying on the complete-data log-likelihood function, which can be expressed as follows:

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{Y}_i,\boldsymbol{U}_i|\boldsymbol{X}_i}(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{x}).$$

It is convenient to decompose this function into three components, each corresponding to a different set of model parameters:

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) = & \sum_{i=1}^{n} \sum_{u=1}^{k} \sum_{t=1}^{T} w_{iu}^{(t)} \log p_{Y_i^{(t)}|U_i^{(t)},\boldsymbol{X}_i^{(t)}}(y^{(t)}|u^{(t)},\boldsymbol{x}^{(t)}) \\ & + \sum_{i=1}^{n} \sum_{u=1}^{k} w_{iu}^{(1)} \log \pi_u + \sum_{i=1}^{n} \sum_{\bar{u}=1}^{k} \sum_{u=1}^{k} z_{i\bar{u}u} \log \pi_{u|\bar{u}}, \end{aligned} \tag{2}$$

where $w_{iu}^{(t)}$ is an indicator variable equal to 1 if subject i is in hidden state u at time t, and $z_{i\bar{u}u} = \sum_{t=2}^{T} w_{i\bar{u}}^{(t-1)} w_{iu}^{(t)}$ denotes the number of times subject i moves from state $\bar{u}$ to state u.

After a suitable initialization of the model parameters, the EM algorithm alternates between the following two steps until convergence:

(*i*) **E-step**: this step computes the conditional expectation of $\ell^*(\boldsymbol{\theta})$ given the observed data and the current parameter estimates. This is achieved by replacing the variables $w_{iu}^{(t)}$ and $z_{i\bar{u}u}$ in (2) with their conditional expected values obtained as follows:

$$\hat{w}_{iu}^{(t)} = q_{U_i^{(t)}|\boldsymbol{X}_i,\boldsymbol{Y}_i}(u|\boldsymbol{x},\boldsymbol{y}), \tag{3}$$

$$\hat{z}_{i\bar{u}u} = \sum_{t=1}^{T} q_{U_i^{(t-1)},U_i^{(t)}|\boldsymbol{X}_i,\boldsymbol{Y}_i}(\bar{u},u|\boldsymbol{x},\boldsymbol{y}). \tag{4}$$

In the previous expressions, $q_{U_i^{(t)}|\boldsymbol{X}_i,\boldsymbol{Y}_i}(u|\boldsymbol{x},\boldsymbol{y})$ denotes the posterior probability of the latent variable at time t given the response and covariate configurations, whereas $q_{U_i^{(t-1)},U_i^{(t)}|\boldsymbol{X}_i,\boldsymbol{Y}_i}(\bar{u},u|\boldsymbol{x},\boldsymbol{y})$ represents the conditional distribution of the latent variables at times $t-1$ and t, given the response and covariate configurations. Note that, in this step, units are not assigned to the latent states, as done in the classification EM algorithm of Celeux and Govaert (1992). Instead, the posterior probabilities of each unit is used to compute the conditional expectation of the complete-data log-likelihood function. The numerical computation of the posterior probabilities in Eqs. (3) and (4) is challenging and requires suitable forward-backward recursions; see Bartolucci et al. (2013, Chapter 5) for further details.

(*ii*) **M-step**: this step maximizes the conditional expectation of $\ell^*(\boldsymbol{\theta})$ and updates parameter estimates. Closed form solutions are available for the initial and transition probabilities:

$$\hat{\pi}_u = \frac{\sum_{i=1}^{n} \hat{w}_{iu}^{(1)}}{n},$$

$$\hat{\pi}_{u|\bar{u}} = \frac{\sum_{i=1}^{n} \hat{z}_{i\bar{u}u}}{\sum_{i=1}^{n}\sum_{v=1}^{k} \hat{z}_{i\bar{u}v}},$$

while solutions for the remaining parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, collected in the vector $\boldsymbol{\theta}_1 = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, are obtained using a Newton-Raphson algorithm. In particular, to reduce the computational burden, we implement a one-step update at each EM iteration. The parameters are then updated as follows:

$$\boldsymbol{\theta}_1^{(h)} = \boldsymbol{\theta}_1^{(h-1)} - \boldsymbol{H}\left(\boldsymbol{\theta}_1^{(h-1)}\right)^{-1} \boldsymbol{s}\left(\boldsymbol{\theta}_1^{(h-1)}\right), \tag{5}$$

where $\boldsymbol{s}(\boldsymbol{\theta}) = \frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}$ and $\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ell^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'}$ are the score vector and the Hessian matrix of the complete-data log-likelihood function, respectively, and h is the iteration counter for the EM algorithm.

To address the multi-modality of the log-likelihood function, parameter initialization typically relies on a multi-start strategy, based on both deterministic and random rules (Bartolucci et al., 2013). The global maximum is then assumed to correspond to the highest value of the log-likelihood function. Regarding convergence of the EM algorithm, two common rules are employed: the algorithm is stopped when the relative change in the log-likelihood function between consecutive iterations falls below a predefined threshold ε_1 and the change in the model parameters between consecutive iterations falls below a predefined threshold ε_2:

$$\frac{\ell(\boldsymbol{\theta}^{(h)}) - \ell(\boldsymbol{\theta}^{(h-1)})}{|\ell(\boldsymbol{\theta}^{(h)})|} < \varepsilon_1 \quad \text{and} \quad \max_s |\theta_s^{(h)} - \theta_s^{(h-1)}| < \varepsilon_2.$$

Standard errors for the parameter estimates can be obtained using the numerical method proposed in Bartolucci and Farcomeni (2009), which involves computing the negative derivative of the score vector at convergence. The score vector is, in turn, obtained as the first derivative of the expected complete data log-likelihood, as suggested in Oakes (1999); see also Pennoni (2014). Concerning model identifiability, establishing global identifiability is generally challenging for HM models. In addition, since the labels of the latent states are arbitrary, the model can be globally identifiable only up to a permutation of the latent states (Bartolucci et al., 2013). However, local identifiability can be assessed by examining the observed information matrix and verifying whether it has full rank (McHugh, 1956; Rothenberg, 1971; Goodman, 1974).

Once the algorithm has converged and the model parameter estimates are available, prediction of the hidden state sequence for each unit is performed through local decoding. This is a separate step from parameter estimation and aims to assign the most likely hidden state to every unit at each time occasion. Specifically, given the estimated posterior probabilities $\hat{w}_{iu}^{(t)}$ obtained at convergence, the predicted state $\hat{u}_i^{(t)}$ for subject i at time t is determined by the maximum-a-posteriori rule:

$$\hat{u}_i^{(t)} = \underset{u=1,\ldots,k}{\operatorname{argmax}}\, \hat{w}_{iu}^{(t)}.$$

The entire sequence of predicted hidden states produced by the local decoding for subject i is denoted as $\hat{\boldsymbol{u}}_i = (\hat{u}_i^{(1)}, \ldots, \hat{u}^{(T)})'$. Considering the posterior probabilities $\hat{w}_{iu}^{(t)}$ in Eq. (3), resulting from the E-step of the estimation algorithm, a weighted mean of the values of $\hat{\alpha}_u$ for each subject i can be computed as

$$\hat{\bar{\alpha}}_i^{(t)} = \sum_{u=1}^{k} \hat{\alpha}_u \hat{w}_{iu}^{(t)}, \tag{6}$$

where $\hat{\bar{\alpha}}_i^{(t)}$ can be referred to as the individual random effects mean.

3. Penalized maximum likelihood estimation

We consider the generalized logit parameterization of the measurement (sub)model, as specified in Eq. (1). In fitted models, when the hidden states are widely separated, subtantial differences in the estimated values of the state-specific parameters α_u may be observed. This may result in (i) greater relevance of one or more states compared with the others, and/or (ii) a reduced estimated contribution of the observed covariates, whose effects may become negligible or statistically insignificant.

In the following sections, we illustrate the challenges posed by separation in HM models through a practical example, present the penalization method along with the corresponding estimation procedure, and describe a cross-validation approach for selecting the strength of the penalization.

3.1. A motivating example

Considering the model introduced in Section 2, we generate data referred to a binary response variable and four covariates, one of which is the lagged response. The regression coefficients are set to $\boldsymbol{\beta} = (1, -1, 1, 1)'$. We assume $k = 3$ latent states, with support points $\boldsymbol{\alpha} = (-20, -5, 5)'$. The initial state probabilities are set to $1/3$ for each state, and the transition probability matrix has diagonal elements equal to 0.750 and off-diagonal elements equal to 0.125. The dataset consists of 250 subjects over 10 time points. The HM model is estimated using the standard maximum likelihood approach described in Section 2.2, and the resulting parameter estimates are reported in Table 1.

The estimated values of α_u are very far apart, resulting in widely separated latent states. This leads to a substantial imbalance in the importance of the different latent states, as confirmed by the allocation of subjects resulting from local decoding. For example, at the initial time occasion, 159 subjects are assigned to the first latent state and 91 to the third, while none are assigned to the second. This contrasts with the true data generation process, in which 88 subjects were assigned to the first state, 71 to the second, and 91 to the third. Moreover, based on the estimated parameters, subjects in the first state consistently have a conditional response probability $\phi^{(t)}_{1|u\boldsymbol{x}}$ equal to 0, whereas those in the third state have a probability equal to 1. This indicates that the covariates have no effect on the estimated conditional probabilities or on the posterior state allocation. It is also worth noting that the estimated values of α_u differ substantially from the true values used in the data generating process. In particular, whereas the true support points included two negative values and one positive value, the estimated values show the opposite pattern, with two positive values and only one negative value. This indicates poor accuracy in parameter estimation. Similar issues arise for the estimated regression coefficients, β_j, which also differ markedly from the values used to generate the data. Moreover, none of the covariates are statistically significant, with all p-values exceeding 0.24.

3.2. Penalization term

To address the problem of latent state separation, we propose to consider the following penalization term, denoted as $\mathcal{A}$:

$$\mathcal{A} = \sum_{u=1}^{k} (\alpha_u - \bar{\alpha})^2 = \sum_{u=1}^{k} \alpha_u^2 - k\bar{\alpha}^2,$$

where $\bar{\alpha} = \frac{1}{k}\sum_{u=1}^{k} \alpha_u$ denotes the mean of the support points. Clearly, the penalty is aimed at limiting the dispersion among the elements of the vector $\boldsymbol{\alpha}$. In matrix notation, we have that

$$\mathcal{A} = \boldsymbol{\alpha}' \boldsymbol{J} \boldsymbol{\alpha},$$

where $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{k}\mathbf{1}\mathbf{1}'$ is the $k \times k$ centering matrix, with $\boldsymbol{I}$ denoting the $k \times k$ identity matrix and $\mathbf{1} = (1, \ldots, 1)'$ a column vector of ones of length k.

The penalization term defined above is applied to both the observed data and the complete data log-likelihood functions, which are modified as follows:

$$\tilde{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \lambda\mathcal{A},$$
$$\tilde{\ell}^*(\boldsymbol{\theta}) = \ell^*(\boldsymbol{\theta}) - \lambda\mathcal{A},$$

where $\lambda \in \mathbb{R}^+$ is a tuning parameter that determines the strength of the penalty term. In this way, by maximizing the penalized log-likelihood function, we seek a balance between model fitting and stability of the estimates of the support points, where stability corresponds to a reduced overall distance among the support points of the hidden states.

As already mentioned, penalized maximum likelihood estimation of the model parameters is performed through a modified version of the EM algorithm, in which the E-step remains unchanged, whereas the M-step requires revising the Newton-Raphson iterative step given in Eq. (5). We denote by $\tilde{\boldsymbol{s}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{H}}(\boldsymbol{\theta})$ the score vector and the Hessian matrix associated with the penalized complete-data log-likelihood, respectively. In particular, the formulas to compute the components corresponding to the vector $\boldsymbol{\alpha}$ are updated as follows:

$$\tilde{\boldsymbol{s}}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{\partial \tilde{\ell}^*(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} = \frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} - 2\lambda \boldsymbol{J}\boldsymbol{\alpha} = \boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) - 2\lambda \boldsymbol{J}\boldsymbol{\alpha},$$
$$\tilde{\boldsymbol{H}}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{\partial^2 \tilde{\ell}^*(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'} = \frac{\partial^2 \ell^*(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'} - 2\lambda \boldsymbol{J} = \boldsymbol{H}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta}) - 2\lambda \boldsymbol{J},$$

Table 1
Maximum likelihood estimates of the HM model parameters with three latent states for data simulated with $\boldsymbol{\beta} = (1, -1, 1, 1)'$ and $\boldsymbol{\alpha} = (-20, -5, 5)'$.

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	−250.126	146.804	457.960	24.142	−80.211	56.686	103.479
Standard error	–	–	–	23.283	69.426	50.905	90.144
p-value	–	–	–	0.300	0.248	0.265	0.251

where $\boldsymbol{s}_{\boldsymbol{\alpha}}$ denotes the sub-vector of $\boldsymbol{s}$ and $\boldsymbol{H}_{\boldsymbol{\alpha\alpha}}$ the sub-matrix of $\boldsymbol{H}$ corresponding to the first and second derivative with respect to $\boldsymbol{\alpha}$. The remaining components of $\tilde{\boldsymbol{s}}$ and $\tilde{\boldsymbol{H}}$ remain unchanged. This is because the penalty term $\mathcal{A}$ depends only on $\boldsymbol{\alpha}$, and therefore its derivative with respect to any other parameter is zero.

3.3. Cross-validation

The tuning parameter λ controls the strength of the penalization in the model. When $\lambda = 0$, no penalty is applied, yielding the same estimates as the standard maximum likelihood method. Conversely, as λ tends to infinity, the support points α_u shrink towards their mean, and the relative importance of unobserved heterogeneity tends to vanish. A reliable method for selecting λ is cross-validation, which is commonly employed in other contexts; see, among others, Kohavi (1995), Smyth (2000), Hastie et al. (2009), and Bates et al. (2023).

In this work, we employ a cross-validation approach to jointly select both the penalization parameter λ and the number of hidden states k. Denoting the full dataset as $\mathcal{D}$, and following the procedure proposed by Smyth (2000) and Bartolucci et al. (2017) for the selection of the number of components in finite mixture models (McLachlan and Peel, 2000) and HM models, respectively, we consider M partitions of the data, $(\mathcal{D}\backslash\mathcal{S}_m, \mathcal{S}_m)_{m=1,\ldots,M}$. For the m-th partition, the model is estimated on the training subset $\mathcal{D}\backslash\mathcal{S}_m$, yielding parameter estimates $\hat{\boldsymbol{\theta}}_{k,\lambda}(\mathcal{D}\backslash\mathcal{S}_m)$. Let $\ell\big(\hat{\boldsymbol{\theta}}_{k,\lambda}(\mathcal{D}\backslash\mathcal{S}_m) \mid \mathcal{S}_m\big)$ denote the log-likelihood function in which the model parameters are estimated on the training data $\mathcal{D}\backslash\mathcal{S}_m$, while the log-likelihood is evaluated on the test data $\mathcal{S}_m$. Finally, the cross-validated likelihood is defined as follows:

$$\ell_{\text{CV}} = \frac{1}{M}\sum_{m=1}^{M} \ell\big(\hat{\boldsymbol{\theta}}_{k,\lambda}(\mathcal{D}\backslash\mathcal{S}_m) \mid \mathcal{S}_m\big),$$

and the optimal value of λ is identified using the one-standard error rule. According to this approach, proposed by Breiman et al. (1984) and recommended, among others, by Hastie et al. (2009) and Krstajic et al. (2014), we select the most regularized model, corresponding to the largest λ, whose cross-validated log-likelihood lies within one standard error of the maximum.

4. Simulation study

A simulation study was conducted to assess the performance of the proposed penalized maximum likelihood approach. We first describe the simulation design (Section 4.1) and then summarize the main results (Sections 4.2, 4.3, and 4.4).

4.1. Simulation design

We consider 48 different simulation scenarios, fixing the number of hidden states at $k = 3$ and varying the following features: sample size ($n = 250, 500$), number of time occasions ($T = 10, 20$), state persistence (high or low), and degree of state separation. For the latter, we define six different configurations for the support points: $\boldsymbol{\alpha}^A = (-3, -1, 0)'$, $\boldsymbol{\alpha}^B = (-15, -5, 0)'$, $\boldsymbol{\alpha}^C = (-45, -15, 0)'$, $\boldsymbol{\alpha}^D = (-4, -1, 1)'$, $\boldsymbol{\alpha}^E = (-20, -5, 5)'$, and $\boldsymbol{\alpha}^F = (-40, -10, 10)'$. For state persistence, two different patterns are considered by specifying the following transition probabilities matrices to represent higher ($\boldsymbol{\Pi}^H$) and lower ($\boldsymbol{\Pi}^L$) levels of persistence:

$$\boldsymbol{\Pi}^H = \begin{bmatrix} 0.900 & 0.050 & 0.050 \\ 0.050 & 0.900 & 0.050 \\ 0.050 & 0.050 & 0.900 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Pi}^L = \begin{bmatrix} 0.750 & 0.125 & 0.125 \\ 0.125 & 0.750 & 0.125 \\ 0.125 & 0.125 & 0.750 \end{bmatrix}.$$

In each scenario, a binary response variable is considered, and four independent standard normally distributed covariates are generated and included in the model specified in Equation (1); the corresponding vector of regression coefficients $\boldsymbol{\beta}$ is set to $(1, -1, 1, 1)'$. For every scenario, 50 random samples are generated, and for each sample the HM model described in Section 2 is estimated using both standard and penalized approaches. To reduce the risk of convergence to local maxima, the estimation algorithm is run 25 times for each sample, using both deterministic and random initialization strategies. The solution yielding the highest likelihood value at convergence is then selected. The following criteria are used to evaluate the results: (*i*) mean squared error (MSE) between the true and estimated model parameters; (*ii*) standard errors of the covariate regression parameters $\boldsymbol{\beta}$; and (*iii*) computational time.

4.2. Performance of the penalized estimation with fixed tuning parameter

In this section, we compare the performance of standard maximum likelihood estimation with that of the proposed penalized approach. As a first step, to assess the effect of regularization, we fix the tuning parameter at $\lambda = 0.01$ across all scenarios. Before

Fig. 1. Percentage variation in the mean squared error between the true and estimated covariate regression coefficient β_j for the penalized estimator ($\lambda = 0.01$), relative to the standard maximum likelihood estimator ($\lambda = 0.00$); the 48 scenarios differ by sample size ($n = 250, 500$), number of time occasions ($T = 10, 20$), state persistence (high or low), and state separation ($\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^F$); each different regression coefficient is represented by a different color.

presenting the results, we note that, when model identifiability was checked using the method introduced in Section 2.2, the observed information matrix was found to be rank-deficient in some samples across 23 of the 48 scenarios considered. This issue affected fewer than 5% of the total samples and occurred only for the estimates obtained without penalization. These findings suggest that the penalized approach may also improve numerical stability and mitigate finite-sample identifiability issues that can arise in latent variable models. The following results exclude these cases to ensure a fair comparison between standard and penalized estimation approaches. To assess the performance of the proposed approach, we first consider the MSE for each covariate regression coefficient. In each scenario, the true model parameter β_j, $j = 1, \ldots, 4$, is compared with the corresponding estimate $\hat{\beta}_{js}$ obtained from sample s, for $s = 1, \ldots, 50$, as follows:

$$\text{MSE}(\beta_j) = \frac{1}{50} \sum_{s=1}^{50} (\beta_j - \hat{\beta}_{js})^2, \quad j = 1, \ldots, 4.$$

Fig. 1 reports the percentage change in $\text{MSE}(\beta_j)$ obtained with penalized maximum likelihood estimation ($\lambda = 0.01$) relative to the standard approach ($\lambda = 0.00$). In summary, we observe that:

- in most scenarios (40 out of 48), the penalized estimation method yields lower MSE values for all regression coefficients β_j. Moreover, all scenarios show a reduction in MSE for at least one coefficient. This suggests that the penalized estimates are, on average, closer to the true values, resulting in improved accuracy;

Table 2
Percentage change in the standard errors of the covariate regression obtained under penalized estimation ($\lambda = 0.01$) relative to the standard approach ($\lambda = 0.00$); the 48 scenarios vary in terms of sample size ($n = 250, 500$), number of time occasions ($T = 10, 20$), state persistence (high or low), and state separation ($\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^F$).

			$\boldsymbol{\alpha}^A$	$\boldsymbol{\alpha}^B$	$\boldsymbol{\alpha}^C$	$\boldsymbol{\alpha}^D$	$\boldsymbol{\alpha}^E$	$\boldsymbol{\alpha}^F$
$\boldsymbol{\Pi}$	T	n	Percentage change in the standard error (%)					
High	10	250	0.76	0.15	2.33	1.86	−90.67	−94.53
		500	4.55	−2.11	0.02	−0.19	−12.27	−95.82
	20	250	−2.83	−4.12	1.05	−0.19	−18.60	−95.73
		500	−0.69	1.72	−2.11	0.01	−6.40	−94.44
Low	10	250	4.02	−2.32	−18.05	−7.53	−76.42	−95.90
		500	0.98	−2.13	−8.75	−11.20	−71.32	−82.19
	20	250	−2.97	8.34	−5.55	−3.98	−28.46	−93.35
		500	−1.31	5.59	2.70	−0.92	−39.20	−76.01

Table 3
Percentage change in computational time obtained under penalized estimation ($\lambda = 0.01$) relative to the standard approach ($\lambda = 0.00$); the 48 scenarios vary in terms of sample size ($n = 250, 500$), number of time occasions ($T = 10, 20$), state persistence (high or low), and state separation ($\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^F$).

			$\boldsymbol{\alpha}^A$	$\boldsymbol{\alpha}^B$	$\boldsymbol{\alpha}^C$	$\boldsymbol{\alpha}^D$	$\boldsymbol{\alpha}^E$	$\boldsymbol{\alpha}^F$
$\boldsymbol{\Pi}$	T	n	Percentage change in computational time (%)					
High	10	250	2.62	−14.31	−2.24	−4.67	−21.46	−52.24
		500	−5.93	−5.90	−14.74	7.58	−7.55	−58.91
	20	250	−2.48	−1.77	19.10	6.10	−15.06	−36.13
		500	−8.29	2.07	3.13	0.17	−7.86	−32.55
Low	10	250	−12.20	−11.63	−0.25	0.74	−22.75	−45.09
		500	20.96	−23.37	4.58	1.33	−16.82	−40.98
	20	250	−7.03	−0.85	−6.99	1.36	−2.38	−22.35
		500	6.16	0.94	5.98	−0.46	−12.01	−24.44

- simulation scenarios corresponding to the fourth and fifth latent state separation patterns ($\boldsymbol{\alpha}^E$ and $\boldsymbol{\alpha}^F$, characterized by highly separated states) exhibit the largest improvements. In these cases, the percentage reduction obtained with penalized estimation often exceeds 90%;
- the penalized approach appears to be less effective under separation patterns characterized by closely spaced states, as well as in settings with a large number of time periods;
- even in the few cases in which the penalized estimation does not improve estimation accuracy, the increase in MSE is very limited, typically remaining below or around 1% and reaching 5% in only one specific case.

A similar behavior is observed when considering the estimated standard errors of the regression coefficients, computed as described in Section 2.2. Table 2 reports the percentage change in these values obtained with the penalized estimation, relative to standard maximum likelihood estimation. Consistently with the previous analyses, penalized estimation leads to a reduction in standard errors in most of the considered scenarios (31 out of 48). As noted above, the proposed approach appears to be less effective under the first state separation levels, especially $\boldsymbol{\alpha}^A$ and $\boldsymbol{\alpha}^B$. In the remaining cases, the percentage reduction is substantial, often reaching very high values and exceeding 90% in scenarios characterized by the greatest separation between latent states ($\boldsymbol{\alpha}^E$ and $\boldsymbol{\alpha}^F$).

Finally, Table 3 reports the percentage variation in computational time for penalized estimation relative to the standard approach. The simulations were carried out on a machine equipped with an Intel Core Ultra 7 155H processor and 32 GB of RAM. Interestingly, the penalized approach often reduces the average computational time, particularly in scenarios corresponding to the fifth and the sixth state separation patterns.

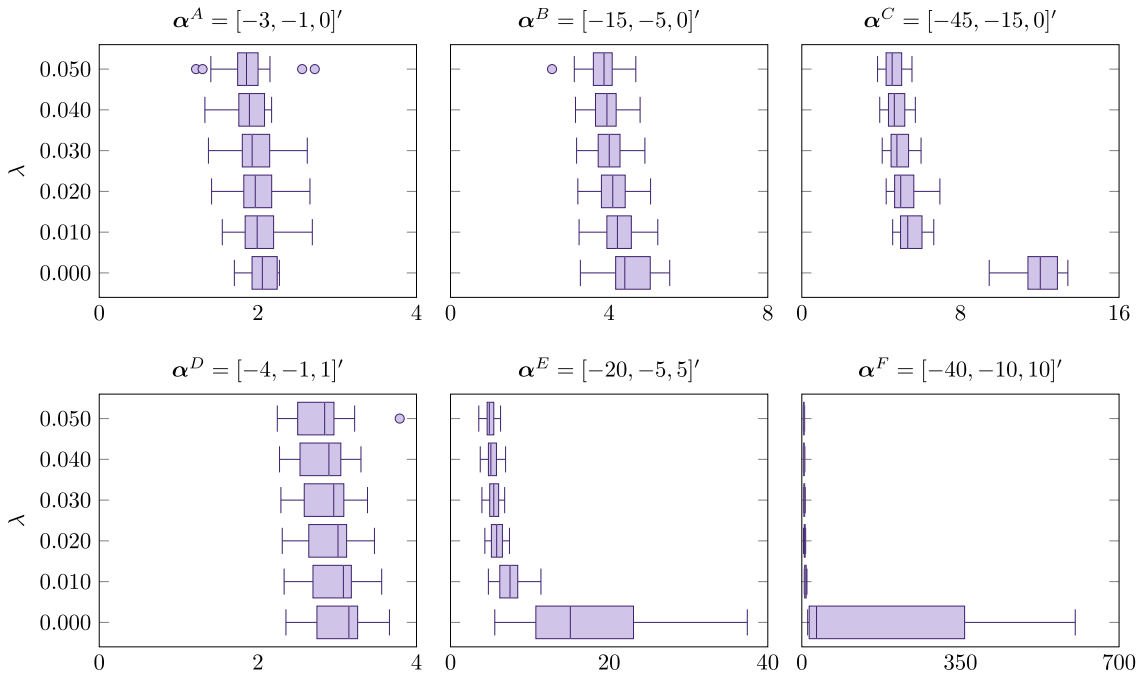


Fig. 2. Variability of the estimated support points across the six latent state separation patterns ($\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^F$), illustrated by boxplots of the estimated standard deviation (x-axis) under penalized estimation ($\lambda \in \{0.01, \ldots, 0.05\}$); the scenarios are defined by a sample size ($n = 250$), number of time occasions ($T = 10$), and low state persistence ($\boldsymbol{\Pi}^L$).

4.3. Comparison across different values of the tuning parameter

We investigate the sensitivity of the proposed estimation method to the choice of the tuning parameter λ by considering six simulation scenarios among the 48 defined in Section 4.1. These scenarios cover all six latent state separation patterns ($\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^F$), while fixing the sample size ($n = 250$), the number of time occasions ($T = 10$), and the state persistence level (low: $\boldsymbol{\Pi}^L$). We compare results obtained with standard and penalized estimation approaches for different values of the tuning parameter λ, ranging from 0 to 0.05 with a step size of 0.01. Fig. 2 shows the variability of the estimated support points, measured by their standard deviation, for each value of λ considered.

The results show that the effect of penalization varies substantially with the degree of latent state separation. For the milder separation patterns ($\boldsymbol{\alpha}^A$, $\boldsymbol{\alpha}^B$, and $\boldsymbol{\alpha}^D$), increasing λ produces only small reductions in the variability of the support points. In contrast, for the more pronounced separation patterns ($\boldsymbol{\alpha}^C$, $\boldsymbol{\alpha}^E$, and $\boldsymbol{\alpha}^F$), variability decreases sharply as λ increases, reflecting the shrinkage induced by the penalty. Since the smallest variability of the support points does not necessarily correspond to the most accurate parameter estimates, a data-driven procedure for selecting λ becomes necessary, as discussed in the next section.

4.4. Performance of cross-validation

In this section, we assess the performance of the penalized estimator when the tuning parameter λ is selected through the cross-validation procedure described in Section 3.3. For each simulated sample, we perform 10-fold cross-validation over the grid of values $\{0.00, 0.01, 0.02, 0.03, 0.04, 0.05\}$ for λ. For each candidate value of λ, the model is estimated on nine folds and evaluated on the remaining one. After rotating the validation fold across all ten partitions, the resulting cross-validated log-likelihood is used to select the optimal value of λ. Fig. 3 evaluates performance in terms of MSE (as defined in Section 3.3) for each covariate regression coefficient β_j, reporting the percentage change in the values obtained using the cross-validated estimation method, relative to the classical maximum likelihood approach.

We observe that, in most cases, the cross-validation approach yields lower MSE values than the standard estimation method, thereby improving the accuracy of the model parameter estimates. This improvement is particularly pronounced in scenarios characterized by the latent state separation patterns $\boldsymbol{\alpha}^E$ and $\boldsymbol{\alpha}^F$, where the reductions in MSE are substantial, consistently exceeding 50% and often approaching 90%. Consistently with the findings from the fixed-λ analysis, the benefits of the penalized approach are less evident in scenarios characterized by lower latent state separation.

Comparing the results in Fig. 1 with those in Fig. 3, we observe that, in scenarios characterized by high state separation, specifically $\boldsymbol{\alpha}^E$ and $\boldsymbol{\alpha}^F$, the results obtained through cross-validation are comparable to those obtained with λ fixed at 0.01. This suggests that, in such extreme cases, the presence of regularization, regardless of the precise tuning, is sufficient to stabilize the estimates. Conversely, in the remaining scenarios, the cross-validation approach outperforms the fixed penalty strategy.

Fig. 3. Percentage variation in the mean squared error between the true and the estimated covariate regression coefficient β_j, computed under the cross-validated penalized estimation relative to the standard approach ($\lambda = 0.00$); the 48 scenarios vary in terms of sample size ($n = 250, 500$), number of time occasions ($T = 10, 20$), state persistence ($\mathbf{\Pi}^H$: high or $\mathbf{\Pi}^L$: low), and state separation ($\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^F$); each different regression coefficient is represented by a different color.

The same conclusions are supported by the estimated standard errors reported in Table 4. For the scenarios with $\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^D$ the percentage decrease of the estimated standard errors is generally moderate, but it exceeds that obtained with the fixed λ approach. Conversely, under higher state separation, the percentage reduction is substantial and very similar to the ones observed in the fixed-λ analysis.

Table 5 shows that, once the tuning parameter has been selected, the penalized model often requires less computational time than the standard approach, although this comparison does not account for the additional cost of the cross-validation procedure.

5. Application

In medical contexts, unobserved heterogeneity is often present, and it is essential to account for it along with relevant covariates that can explain the observed heterogeneity among patients (Schlattmann, 2009). To address this need, random effects and finite mixture models are often employed. However, when longitudinal data are available, the HM model may provide a more informative approach than generalized linear models (Liang and Zeger, 1986), which can be regarded as the restricted case of a single latent state. Indeed, the HM framework makes possible to disentangling time-dependent variability and identifying unobserved groups of patients with different progressions. In particular, clustering patients may be very important in precision medicine to enhance prediction accuracy for tailored interventions.

Table 4
Percentage change in the standard errors of the covariates regression coefficients obtained under the cross-validated penalized estimation relative to the standard approach ($\lambda = 0.00$); the 48 scenarios vary in terms of sample size ($n = 250, 500$), number of time occasions ($T = 10, 20$), state persistence ($\mathbf{\Pi}^H$ high or $\mathbf{\Pi}^L$ low), and state separation ($\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^F$).

			$\boldsymbol{\alpha}^A$	$\boldsymbol{\alpha}^B$	$\boldsymbol{\alpha}^C$	$\boldsymbol{\alpha}^D$	$\boldsymbol{\alpha}^E$	$\boldsymbol{\alpha}^F$
$\mathbf{\Pi}$	T	n	Percentage change in the standard error (%)					
High	10	250	−3.26	1.80	−16.82	−1.83	−93.18	−96.21
		500	−3.93	−5.19	0.02	−1.24	−19.71	−96.72
	20	250	−1.64	−6.61	−0.88	−0.77	−32.32	−96.10
		500	−1.09	−2.82	−2.63	−0.42	−11.38	−94.66
Low	10	250	2.79	−6.58	−17.22	5.29	−74.37	−96.19
		500	−39.17	−13.04	5.84	−6.48	−79.40	−84.80
	20	250	−13.26	−7.72	−7.40	−13.14	−41.91	−95.31
		500	2.34	−2.54	6.92	1.82	−49.93	−82.64

Table 5
Percentage change in computational time obtained under the cross-validated penalized estimation relative to the standard approach ($\lambda = 0.00$); the 48 scenarios vary in terms of sample size ($n = 250, 500$), number of time occasions ($T = 10, 20$), state persistence ($\mathbf{\Pi}^H$ high or $\mathbf{\Pi}^L$ low), and state separation ($\boldsymbol{\alpha}^A, \ldots, \boldsymbol{\alpha}^F$).

			$\boldsymbol{\alpha}^A$	$\boldsymbol{\alpha}^B$	$\boldsymbol{\alpha}^C$	$\boldsymbol{\alpha}^D$	$\boldsymbol{\alpha}^E$	$\boldsymbol{\alpha}^F$
$\mathbf{\Pi}$	T	n	Percentage change in computational time (%)					
High	10	250	−11.08	−12.72	−16.91	−2.56	−33.83	−49.78
		500	−4.34	−6.07	−14.74	−6.65	−1.30	−68.48
	20	250	−6.91	6.01	23.72	1.36	−24.27	−28.07
		500	−9.68	−17.50	−3.72	−0.52	−10.98	−14.27
Low	10	250	−24.36	−18.25	−0.99	−19.02	−23.51	−45.89
		500	−11.41	−17.16	−10.27	−2.34	−25.82	−46.80
	20	250	3.47	−14.82	−19.26	−1.63	−12.92	−33.87
		500	1.05	−5.19	−5.15	2.64	−14.97	−38.99

In the following, we apply the proposed HM model to medical longitudinal data collected at the Anesthesiology and Reanimation Department of Akdeniz University Hospital, as presented and described in Aktas et al. (2014). The data are freely available at https://peerj.com/articles/648/. We first describe the data and then interpret the results obtained from the inferential procedures illustrated in Section 3.

5.1. Data description

The data include information on patients undergoing spinal anesthesia during a surgery, with records for $n = 417$ patients collected from January 2008 to January 2011. Following Aktas et al. (2014), we consider only patients aged 17 and over, resulting in $n = 375$ individuals. Various complications may arise during spinal anesthesia. Among these, hypotension (Sharma et al., 1997; Somboonviboon et al., 2008) involves a decrease in systolic blood pressure that may potentially lead to death (Sanborn et al., 1996). The data were collected to monitor the evolution of this phenomenon during surgery and ensure patient safety. Hypotensive status can be defined according to various criteria. In the study by Aktas et al. (2014), which follows one of the most common criteria (Klöhr et al., 2010), the binary response variable is defined as $y = 1$ (hypotensive status) if the systolic blood pressure (SBP) is below 100 mmHg or below 80% of baseline SBP, and as $y = 0$ (non-hypotensive status) otherwise. Measurements were collected over eight occasions, at equally spaced intervals across a 40 min period.

Individual-specific characteristics were also recorded, including time-fixed covariates such as gender, age, type of surgical hospital unit (general surgery, urology, obstetrics and gynecology), patient position during surgery (lithotomy, supine), electrocardiography (ECG) status (normal, abnormal), and medication doses in the blood (Marcain-heavy, Chirocaine, Fentanyl and Midazolam). Time-varying covariates include diastolic blood pressure (DBP) and patient pulse rate.

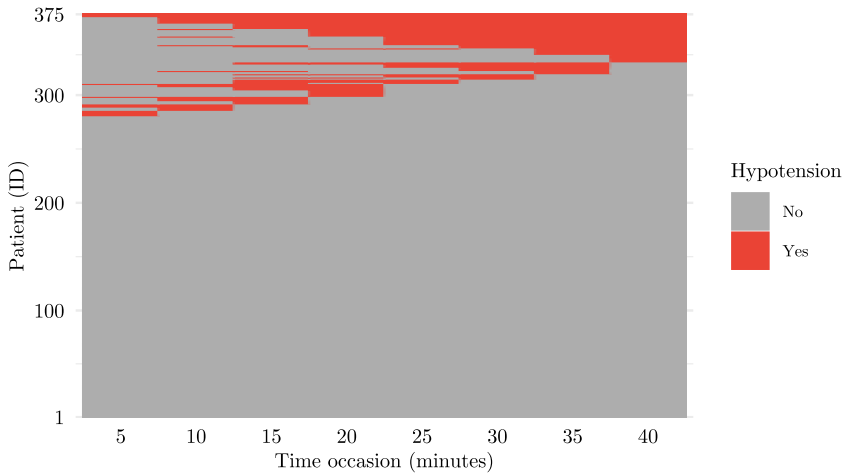


Fig. 4. Occurrences of hypotension during spinal anesthesia for each patient; measurements are taken at five-minute intervals, starting five minutes after the beginning of surgery.

Table 6
Descriptive statistics of anesthetic drugs dose.

Anesthetic drug	Frequency	Mean ± SD
Chirocaine	143	5.293 ± 7.895
Fentanyl	116	0.033 ± 0.059
Marcain-heavy	229	7.115 ± 6.181
Midazolam	202	0.821 ± 1.003

The lasagna plot in Fig. 4 illustrates the observed hypotension occurrences across the eight measurement occasions. Approximately 25% ($n = 94$) of the patients experienced at least one hypotensive episode during this period. These findings are consistent with the literature, which reports an incidence of hypotension ranging from 15% to 33% of patient cases; see, among others, Carpenter et al. (1992), Hartmann et al. (2002), Lin et al. (2008). The average age of patients was approximately 49 years old. Out of the 375 individuals, 56% ($n = 210$) were male, and the remaining 44% ($n = 165$) female. Surgeries were conducted in three different hospital units: 38% in obstetrics and gynecology, 44% ($n = 165$) in urology, and the remaining 18% ($n = 66$) in general surgery. Overall, 58% ($n = 218$) of the procedures were performed with patients in the supine position, while 42% ($n = 157$) were in the lithotomy position. The ECG status was normal for 98% of patients ($n = 366$), while 2% ($n = 9$) presented abnormal readings.

Summary statistics of the doses administered for each drug during anesthesia are presented in Table 6. Note that each patient may receive more than one medication during surgery.

5.2. Results

The HM model is estimated with covariates, including the lagged hypotension status, in order to account for state dependence and relax the conditional independence assumption.

Cross-validation, as described in Section 3, is employed to jointly select the optimal number of hidden states k and the penalty parameter λ. Specifically, we consider $k = 1, \dots, 4$ and a grid of λ values ranging from 0 to 0.05 with a step size of 0.005. We also test larger values, $\lambda = 1, 5$; however, these results are not reported, as they yielded suboptimal performance due to excessive smoothing. The results, in terms of cross-validated log-likelihood values, are shown in the heat map in Fig. 5. The best value, equal to −48.973, is achieved by the model with $k = 3$ hidden states, 23 parameters, and penalization parameter $\lambda = 0.02$.

Table 7 presents the estimates of the parameters affecting the conditional distribution of the response variables given the latent process, together with their estimated standard errors, p-values, and significance levels.

Older individuals exhibit higher log-odds of being diagnosed with hypotension than younger individuals. Similarly, being female has a significant positive effect on the conditional probability of experiencing hypotension, given the hidden state. DBP has a significant negative effect on the log-odds of hypotension, meaning that a lower DBP is associated with higher odds of hypotension. The lagged response has a significant positive effect on hypotension, indicating serial correlation. Additionally, the supine position is associated with higher log-odds of hypotension than the lithotomy position. Regarding the anesthetics used, Midazolam has a significant positive effect, indicating that higher concentration of this drug in the blood is associated with increased odds of experiencing hypotension during surgery. For the other drugs, the estimated coefficients are not statistically significant. It is important to note that the effect of some of these drugs on hypotension is still the subject of ongoing research, as discussed in the recent review by Paliwal et al. (2024).

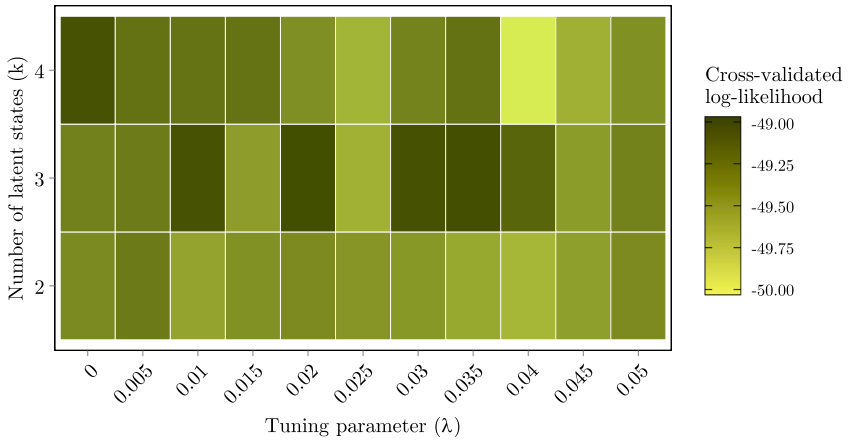


Fig. 5. Heat map of the cross-validated log-likelihood values ($\hat{\ell}_{CV}$) under the HM model, estimated for a number of hidden states (k) ranging from 2 and 4, and for a grid of λ values ranging from 0 to 0.05 with a step size of 0.005; to preserve the resolution of the color scale, results for $k = 1$ and for larger penalty values ($\lambda = 1, 5$) are not shown due to their particularly low values.

Table 7
Estimates of the regression parameters affecting the conditional response probabilities, together with their standard errors and corresponding p-values under the HM model with $k = 3$ hidden states and $\lambda = 0.02$. Significance is indicated at the †10%, *5%, and **1% levels.

Covariate	$\hat{\beta}$	$s.e.$	p-value
Age (year)	+0.036*	0.019	0.042
Chirocaine	−0.049	0.036	0.180
DBP	−0.166**	0.022	0.000
ECG (Normal)	−0.858	0.893	0.337
Fentanyl	+1.210	2.350	0.608
Gender (Female)	+1.526*	0.752	0.042
Hypotension ($t-1$)	+2.678**	0.259	0.000
Marcain-heavy	−0.049	0.040	0.220
Midazolam	+0.242*	0.115	0.036
Operation (General surgery)	−0.058	0.897	0.949
Operation (Urology)	+0.374	1.073	0.727
Position (Supine)	+0.797	0.508	0.116
Pulse rate	−0.002	0.011	0.858

Table 8
Support points, initial probabilities, and transition probabilities values under the HM model with $k = 3$ hidden states and $\lambda = 0.02$.

Estimate	u		
	1	2	3
$\hat{\alpha}_u$	−0.821	3.129	7.260
$\hat{\pi}_u$	0.816	0.161	0.023
$\hat{\pi}_{u\|1}$	0.991	0.009	0.000
$\hat{\pi}_{u\|2}$	0.030	0.954	0.014
$\hat{\pi}_{u\|3}$	0.108	0.000	0.892

Table 8 shows the estimated support points for the hidden states α_u, as defined in Eq. (1), together with the initial and transition probabilities. The estimated support points provide clinically meaningful labels for the patients subpopulations corresponding to the hidden states identified by the HM model, suggesting a natural ordering. The 1st subpopulation corresponds to patients with the lowest propensity for hypotension, where as the 3rd corresponds to those with the highest propensity. Fig. 6 depicts the trend of the estimated conditional probability of hypotension for each subpopulation over time. Patients in the 1st subpopulation show a null probability of hypotension throughout the surgery. Patients in the 2nd subpopulation show a low probability of hypotension, ranging approximately from 0.10 to 0.20. In contrast, patients in the 3rd subpopulation exhibit a very high probability of hypotension during

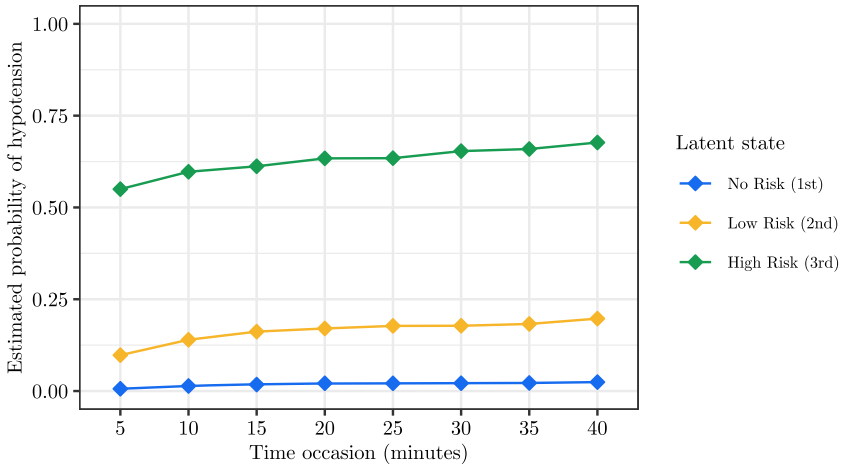


Fig. 6. Estimated probability of hypotension under the HM model with $k = 3$ hidden states and $\lambda = 0.02$, for each state u at each time occasion.

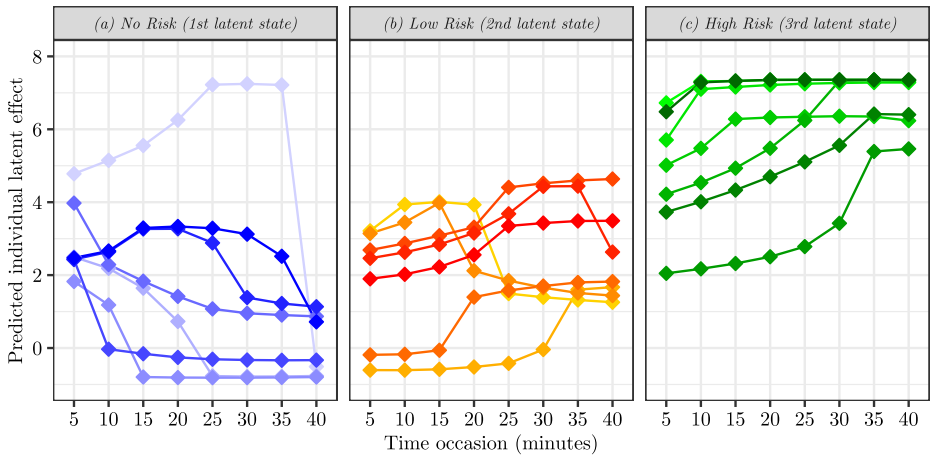


Fig. 7. Estimated average individual latent effect ($\hat{\tilde{\alpha}}$) at each time occasion for selected individuals, classified at the final time point as follows: the left panel (*a*) refers to seven patients classified as No Risk (1st state), middle panel (*b*) to seven patients classified as Low Risk (2nd state), and the right panel (*c*) to seven patients classified as High Risk (3rd state).

surgery, ranging from 0.54 to 0.68, with an increasing trend over time. Therefore, the groups can be labeled as follows: the 1st subpopulation as “No Risk”, the 2nd as “Low Risk”, and the 3rd as “High Risk”.

The estimated initial probabilities reported in Table 8 suggest that, at the beginning of surgery, approximately 82% of patients belong to the 1st subpopulation (No Risk), 16% to the 2nd subpopulation (Low Risk), and 2% to the 3rd subpopulation (High Risk). The estimated transition probabilities indicate high persistence within the 1st and 2nd subpopulations, with about 1.4% of patients transitioning from the low-risk (2nd subpopulation) to the high-risk subgroup (3rd subpopulation). In contrast, the high-risk subgroup shows lower persistence, with 10.8% of patients estimated to transition to the no-risk group.

Patients are dynamically assigned to the hidden states using local decoding, as described in Section 2, based on the estimated posterior probabilities. At each time occasion, every patient is assigned to the subgroup with the highest posterior membership probability. According to this estimated posterior distribution, approximately 86.7–87.6% of individuals are classified into the 1st subgroup (No Risk) over the considered time period, 10.9–11.5% in the 2nd subgroup (Low Risk), and 1.6-2.1% in the 3rd subgroup (High Risk). The model may assist physicians in identifying patients at high risk of hypotension during the early stages of surgery, enabling timely preventive measures to improve patient safety.

Fig. 7 shows the estimated values $\hat{\tilde{\alpha}}_i^{(t)}$ for selected individuals. Specifically, the left panel (*a*) refers to seven patients clustered in the 1st subgroup (No Risk) at the last measurement occasion. These patients show a generally decreasing trend in the estimated values over time, indicating a reducing risk of hypotension. The middle panel (*b*) concerns other seven patients clustered in the 2nd subgroup (Low Risk). Their trajectories are more heterogeneous over time, with some patients experiencing a slight increase and others a slight decrease in the risk during surgery. The right panel (*c*) refers to seven patients clustered in the 3rd subgroup (High Risk) at the last measurement occasion, indicating that these patients tend to have an increasing risk of hypotension during the observed period.

It is worth noting that the HM model proposed here differs substantially from both the generalized estimating equation approach (Liang and Zeger, 1986) and the generalized linear mixed model (Fitzmaurice et al., 2012) used to analyze the same data in Aktas et al. (2014). The former describes changes in the population mean, whereas the latter analyzes changes in individual response means. Although these alternative models use a number of parameters comparable to that of the proposed model, both require the estimation of a breakpoint at the 20 min time point to better capture time dynamics. The estimated covariate effects are larger in magnitude under the HM model; moreover, gender, patient position during surgery, and the use of Midazolam are not statistically significant under the generalized estimating equation and mixed model approaches. The HM model accounts for marginal effects, while, within the proposed formulation, individual heterogeneity is captured by the random cluster effects. This makes it possible, for example, to disentangle the pattern observed in Fig. 7, which, is instead captured through a breakpoint in the other two models.

6. Conclusions

In this study, we introduce a new estimation method for hidden Markov (HM) models with covariates that directly affect the response variables. The model accounts for unobserved heterogeneity and serial dependence by including the lagged response among the covariates. It is particularly useful for dynamically classifying subjects into groups based on the estimated posterior probabilities, while also explaining the evolution of the phenomenon of interest over time. We propose a penalized maximum likelihood estimation approach implemented through the expectation-maximization algorithm. The penalty primarily prevents excessively larger estimates of the support points related to the latent states, leading to more accurate parameter estimates. In turn, given the iterative alternation between the E-step and M-step of the algorithm, this also results in more reliable posterior probabilities. Another important implication concerns model identifiability, which is supported by the increased stability of the penalized estimates. In addition, to the best of our knowledge, this is the first work to propose a cross-validation approach for HM models to jointly select k and λ within this penalized HM formulation. This model selection strategy can be viewed as an alternative to information criteria, which are commonly used when the number of subpopulations is not known a priori.

The proposal is validated through an extensive simulation study, in which we assess the effectiveness of the procedure under different scenarios. The simulation results show improved accuracy in parameter estimation and, for the fitted model after selection of the tuning parameter, reduced computational time. However, the overall computational burden increases when cross-validation over a grid of values for k and λ is required. Selection of the optimal tuning parameter λ through the cross-validation procedure is particularly useful in scenarios characterized by moderate latent state separation. Conversely, when separation is extreme, penalization ensures substantial improvements regardless of the specific value of λ.

The empirical illustration shows that the model can detect the effects of significant covariates, such as gender and the type of drug used during surgery, on hypotension. It is also effective in classifying patients into groups with increasing severity, making it particularly valuable for real-time monitoring of high-risk patients during surgery and, therefore, as an early warning system. In addition, we show that the employed model differs from alternative models commonly used for the analysis of longitudinal data, particularly because it allows for straightforward treatment of serial dependence and the identification of patient clusters with different levels of hypotension risk and distinct trajectories over time.

The proposed penalty is related to ridge-type regularization, although it differs from the standard ridge penalty because it acts on the deviations of the support points from their mean rather than on their absolute magnitude. This choice preserves the overall location of the latent support points while discouraging excessive separation among states. From a Bayesian perspective, the penalty can be interpreted as imposing a Gaussian prior on the centered latent support points; see Tibshirani (1996) for further details. An alternative approach could involve a lasso-type penalization, which corresponds to the use of Laplace priors in the Bayesian estimation.

Some limitations of the proposed approach have to be acknowledged. First, the present formulation is tailored to a univariate binary response, with covariates affecting the measurement model. A possible extension would be to assess the penalty when a categorical response variable is considered. Second, the cross-validation procedure can be computationally demanding, because several models must be estimated over a grid of values for k and λ. Finally, the proposed penalty is tailored to latent-state separation only, and additional penalties on the regression coefficients may be needed in high-dimensional settings.

Several extensions could also be explored along this research line. From a methodological perspective, further investigation is needed to understand the circumstances under which large estimates of the support points of latent states occur. The model could be extended to accommodate multivariate response variables of mixed types, including both continuous and categorical variables, handle missing values, and account for lagged dependencies of second or higher order. From a computational perspective, a key priority is to improve the scalability of the estimation procedure. The computational time required to estimate the HM model with covariates depends on the number of latent states and covariates, and it can be substantial, particularly when cross-validation is needed to tune the parameter λ. To mitigate this issue, future work could focus on implementing the core expectation-maximization algorithm in C++ (Stroustrup, 2013) or Fortran, with the aim of incorporating the proposed method into the `LMest` package (Pennoni et al., 2025). This is expected to substantially reduce execution time compared to the current implementation. Furthermore, since the grid search for λ requires estimating many HM models independently, distributing these tasks across multiple cores through parallel computation would significantly reduce the overall computational burden. An alternative to exhaustive grid search could involve the use of genetic algorithms, where the fitness function is defined as the cross-validated log-likelihood. Finally, for high-dimensional datasets, integrating dimensionality reduction techniques will be essential for efficiently handling large-scale applications.

Acknowledgments

The authors acknowledge the financial support from the grant “Hidden Markov Models for Early Warning Systems” of Ministero dell’Università e della Ricerca (PRIN 2022TZEXKF) funded by the European Union - Next Generation EU, Mission 4, Component 2, CUP J53D23004990006.

References

Agresti, A., 2002. Categorical Data Analysis, 2nd Edition. John Wiley & Sons, Hoboken, USA.
Aktas, S.A., Coskunfirat, N., Saka, O., 2014. Comparison of predictor approaches for longitudinal binary outcomes: application to anesthesiology data. PeerJ 2, 1–15.
Alexandrovich, G., 2016. Penalized maximum likelihood estimation for Gaussian hidden Markov models. Commun. Stat. – Theory Methods 45, 6133–6148.
Bartolucci, F., Chiaromonte, F., Kuruppumullage Don, P., Lindsay, B.G., 2017. Composite likelihood inference in a discrete latent variable model for two-way “clustering-by-segmentation” problems. J. Comput. Graph. Stat. 26, 388–402.
Bartolucci, F., Farcomeni, A., 2009. 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.
Bartolucci, F., Farcomeni, A., Pennoni, F., 2013. Latent Markov Models for Longitudinal Data. Chapman and Hall/CRC, Boca Raton, USA.
Bartolucci, F., Farcomeni, A., Pennoni, F., 2014. Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. Test 23, 433–465.
Bartolucci, F., Pennoni, F., 2007. A class of latent Markov models for capture–recapture data allowing for time, heterogeneity, and behavior effects. Biometrics 63, 568–578.
Bates, D., Hastie, T., Tibshirani, R., 2023. Cross-validation: what does it estimate and how well does it do it? J. Mach. Learn. Res. 24, 1234–1256.
Baum, L.E., Petrie, T., 1966. Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. 37, 1554–1563.
Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Chapman and Hall/CRC, New York, USA.
Cagnone, S., Moustaki, I., Vasdekis, V., 2009. Latent variable models for multivariate longitudinal ordinal responses. Br. J. Math. Stat. Psychol. 62, 401–415.
Carpenter, R.L., Caplan, R.A., Brown, D.L., Stephenson, C., Wu, R., 1992. Incidence and risk factors for side effects of spinal anesthesia. Anesthesiology 76, 906–916.
Celeux, G., Govaert, G., 1992. A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data Anal. 14, 315–332.
Clark, R.G., Blanchard, W., Hui, F.K.C., Tian, R., Woods, H., 2023. Dealing with complete separation and quasi-complete separation in logistic regression for linguistic data. Res. Methods Appl. Linguist. 2, 1–10.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B 39, 1–38.
Farcomeni, A., 2017. Penalized estimation in latent Markov models, with application to monitoring serum calcium levels in end-stage kidney insufficiency. Biom. J. 59, 1035–1046.
Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2012. Applied Longitudinal Analysis. John Wiley & Sons, Hoboken, USA.
Good, I.J., Gaskins, R.A., 1971. Nonparametric roughness penalties for probability densities. Biometrika 58, 255–277.
Goodman, L.A., 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61, 215–231.
Hartmann, B., Junger, A., Klasen, J., Benson, M., Jost, A., Banzhaf, A., Hempelmann, G., 2002. The incidence and risk factors for hypotension after spinal anesthesia induction: an analysis with automated data collection. Anesth. Analg. 94, 1521–1529.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer, New York, USA.
Hsiao, C., 2003. Analysis of Panel Data, 2nd Edition. Cambridge University Press, Cambridge, United Kingdom.
Hung, Y., Wang, Y., Zarnitsyna, V., Zhu, C., Wu, C.F.J., 2013. Hidden Markov models with applications in cell adhesion experiments. J. Am. Stat. Assoc. 108, 1469–1479.
Klöhr, S., Roth, R., Hofmann, T., Rossaint, R., Heesen, M., 2010. Definitions of hypotension after spinal anaesthesia for caesarean section: literature search and application to parturients. Acta Anaesthesiol. Scand. 54, 909–921.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Vol. 2. Montreal, Canada, pp. 1137–1145.
Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminformatics 6, 1–15.
Langrock, R., Kneib, T., Sohn, A., De Ruiter, S.L., 2015. Nonparametric inference in hidden Markov models using P-splines. Biometrics 71, 520–528.
Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.
Lin, C.-S., Chiu, J.-S., Hsieh, M.-H., Mok, M.S., Li, Y.-C., Chiu, H.-W., 2008. Predicting hypotensive episodes during spinal anesthesia with the application of artificial neural networks. Comput. Methods Programs Biomed. 92, 193–197.
Lin, Y., Song, X., 2022. Order selection for regression-based hidden Markov model. J. Multivar. Anal. 192, 1–20.
Mansournia, M.A., Geroldinger, A., Greenland, S., Heinze, G., 2017. Separation in logistic regression: causes, consequences, and control. Am. J. Epidemiol. 187, 864–870.
Maruotti, A., Punzo, A., 2017. Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. Comput. Stat. Data Anal. 113, 475–496.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd Edition. Chapman and Hall, CRC, London, United Kingdom.
McCulloch, C.E., Searle, S.R., Neuhaus, J.M., 2008. Generalized Linear Mixed Models. John Wiley & Sons, New York, USA.
McHugh, R.B., 1956. Efficient estimation and local identification in latent class analysis. Psychometrika 21, 331–347.
McLachlan, G., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, New York, USA.
Mor, B., Garhwal, S., Kumar, A., 2021. A systematic review of hidden Markov models and their applications. Arch. Comput. Methods Eng. 28, 1429–1448.
Oakes, D., 1999. Direct calculation of the information matrix via the EM algorithm. J. R. Stat. Soc. B 61, 479–482.
Ötting, M., Groll, A., 2022. A regularized hidden Markov model for analyzing the ‘hot shoe’ in football. Stat. Model. 22, 546–565.
Paliwal, N., Kokate, M.V., Deshpande, N.A., Khan, I.A., 2024. Spinal anaesthesia using hypobaric drugs: a review of current evidence. Cureus 16, 1–8.
Pennoni, F., 2014. Issues on the Estimation of Latent Variable and Latent Class Models. Scholars’ Press, Saarbrücken, Germany.
Pennoni, F., Pandolfi, S., Bartolucci, F., 2025. LMest: An R package for estimating generalized latent Markov models. R J. 16, 74–101.
Radwan, M.A., O’Carroll, L., McCaul, C.L., 2024. Total spinal anaesthesia following obstetric neuraxial blockade: a narrative review. Int. J. Obstet. Anesth. 59, 1–12.
Rothenberg, T.J., 1971. Identification in parametric models. Econometrica 39, 577–591.
Sanborn, K.V., Castro, J., Kuroda, M., Thys, D.M., 1996. Detection of intraoperative incidents by electronic scanning of computerized anesthesia records: comparison with voluntary reporting. Anesthesiology 85, 977–987.
Schlattmann, P., 2009. Medical Applications of Finite Mixture Models. Springer-Verlag, Heidelberg, Germany.
Scott, D.W., Tapia, R.A., Thompson, J.R., 1978. Multivariate density estimation by discrete maximum penalized likelihood methods. In: Graphical Representation of Multivariate Data. Elsevier, pp. 169–182.
Sharma, S.K., Gajraj, N.M., Sidawi, J.E., 1997. Prevention of hypotension during spinal anesthesia: a comparison of intravascular administration of hetastarch versus lactated ringer’s solution. Anesth. Analg. 84, 111–114.
Silverman, B.W., 1982. On the estimation of a probability density function by the maximum penalized likelihood method. Ann. Stat. 10, 795–810.
Smyth, P., 2000. Model selection for probabilistic clustering using cross-validated likelihood. Stat. Comput. 9, 63–72.
Somboonviboon, W., Kyokong, O., Charuluxananan, S., Narasethakamol, A., 2008. Incidence and risk factors of hypotension and bradycardia after spinal anesthesia for cesarean section. J. Med. Assoc. Thail. 91, 181–187.

Stroustrup, B., 2013. The C++ Programming Language, 4th Edition. Addison-Wesley, Boston, USA.
R Core Team, 2025. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B 58, 267–288.
Visser, I., Speekenbrink, M., 2022. Mixture and Hidden Markov Models with R. Springer, Cham, Switzerland.
Welch, L.R., 2003. Hidden Markov models and the Baum-Welch algorithm. IEEE Inform. Theory Soc. Newsl. 50, 10–13.