



Tempered expectation-maximization algorithm for the estimation of discrete latent variable models

Luca Brusa¹ · Francesco Bartolucci² · Fulvia Pennoni³

Received: 12 January 2022 / Accepted: 10 August 2022

© The Author(s) 2022

Abstract

Maximum likelihood estimation of discrete latent variable (DLV) models is usually performed by the expectation-maximization (EM) algorithm. A well-known drawback is related to the multimodality of the log-likelihood function so that the estimation algorithm can converge to a local maximum, not corresponding to the global one. We propose a tempered EM algorithm to explore the parameter space adequately for two main classes of DLV models, namely latent class and hidden Markov. We compare the proposal with the standard EM algorithm by an extensive Monte Carlo simulation study, evaluating both the ability to reach the global maximum and the computational time. We show the results of the analysis of discrete and continuous cross-sectional and longitudinal data referring to some applications of interest. All the results provide supporting evidence that the proposal outperforms the standard EM algorithm, and it significantly improves the chance to reach the global maximum. The advantage is relevant even considering the overall computing time.

Keywords Annealing · Global maximum · Hidden Markov model · Latent class model · Local maxima

✉ Luca Brusa
luca.brusa@unimib.it

Francesco Bartolucci
francesco.bartolucci@unipg.it

Fulvia Pennoni
fulvia.pennoni@unimib.it

¹ Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

² Department of Economics, University of Perugia, Perugia, Italy

³ Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

1 Introduction

A latent variable model is a statistical model in which the distribution of the response variables is affected by one or more variables that are not directly observable. Here, we consider two special classes of discrete latent variable (DLV) models (Bartolucci et al. 2022) that are frequently employed to analyze continuous and categorical response variables.

The latent class (LC) model (Lazarsfeld and Henry 1968; Goodman 1974; Lindsay et al. 1991) assumes individual-specific latent variables having a discrete distribution with a finite number of support points. The hidden (or latent) Markov (HM) model (Zucchini and Guttorp 1991; Bartolucci et al. 2013; Zucchini et al. 2016) represents a generalization of the LC model to the case of longitudinal data and the latent process is frequently assumed as a first-order Markov chain. Both models are used as model-based clustering methods, and in particular, the HM model allows a dynamic clustering where each unit may move between clusters across time.

Maximum likelihood estimation (MLE) of DLV models is usually performed by using the expectation-maximization (EM) algorithm (Baum et al. 1970; Dempster et al. 1977; McLachlan and Krishnan 2008). This approach is straightforward to implement, and it is available in many software packages; among others, we mention `MultilLCIRT` (Bartolucci et al. 2014) and `LMest` (Bartolucci et al. 2017) in the R software (R Core Team 2022) for the estimation of LC and HM models, respectively.

A particular drawback of MLE is related to the multimodality of the log-likelihood function which is especially observed with the DLV models. Consequently, the EM algorithm could converge to a local maximum, not corresponding to the global one. Multi-start strategies employing both deterministic and random rules to initialize the model parameters are generally adopted. Although this approach encourages a more accurate exploration of the parameter space, it is computationally intensive and does not ensure that the global optimum is reached. For an overview of different initialization strategies, some of which are based on a preliminary cluster analysis (Everitt et al. 2011), see, among others, Maruotti and Punzo (2021).

Tempering and annealing (Sambridge 2014) constitute a broad family of optimization methods; by means of a parameter known as temperature, they allow us to re-scale the target function and monitor the prominence of all possible maxima. In particular, these procedures are gradually attracted towards the global optimum by accurately defining a sequence of temperature values. The alternation of high and low values of the temperature allows us to deal with two opposite but fundamental issues: on one side, the algorithm is led to explore broad areas of the parameter space, thus escaping local sub-optimal modes (high temperatures); on the other side, the algorithm is able to perform a sharp optimization of the target function in a small area of the parameter space (low temperatures).

The following different tempering methods are defined according to the choices of temperature sequences. Simulated annealing (Kirkpatrick et al. 1983) makes use of a strictly decreasing temperature sequence: the initial temperature is sufficiently high so that the re-scaled function is relatively flat, and it decreases at each step, gradually restoring the original function. Simulated tempering (Geyer and Thompson 1995) assumes that the temperature may either increase or decrease according to a stochastic

rule: a new proposed temperature level may be accepted or rejected according to a specific probability, and the process describing the temperature evolution follows a Markov chain. Parallel tempering (Geyer 1991; Falcioni and Deem 1999; Earl and Deem 2005) assumes an ensemble of Markov chains across all levels of the temperature sequence: at specified intervals, a swap between a pair of neighboring chains is proposed and accepted or rejected according to a certain probability.

Tempering techniques are employed, among others, in Barbu and Zhu (2013) and Robert et al. (2018) for simulating from complex multimodal statistical distributions by means of Markov chain Monte Carlo methods (Metropolis et al. 1953; Hastings 1970). On the other hand, the use of these procedures within the EM algorithm is quite scarce. Hofmann (1999) proposed tempering techniques for the EM algorithm in the context of probabilistic latent semantic analysis. For what concerns finite Gaussian mixture models, recently, Lartigue et al. (2022) proposed a general class of deterministic approximated versions of the EM algorithm following previous proposals in Yuille et al. (1994), Ueda and Nakano (1998), and Zhou and Lange (2010).

In the following, dealing with DLV models, we propose a general approach. In particular, we explicitly focus on LC and HM models because these are among the most utilized LDV models in data analysis. However, the proposal can easily be adapted to the aforementioned finite mixture models and to other DLV models. We explore two different temperature sequences, including a non-monotone one, also evaluating the computational time efficiency. Up to our knowledge, for the first time, we deal with the problem of temperature sequence tuning, inspecting the performance of the tempered EM (T-EM) algorithm with both optimally tuned and fixed temperature sequences. Finally, we show the behavior of the algorithm for the selection of the optimal model. The implemented code for the proposal is written for the open source software R (R Core Team 2022). It is based on some functions of the package `LMest` (Bartolucci et al. 2017), and it is available at the following link in the GitHub repository: <https://github.com/LB1304/T-EM>.

The remainder of the paper is organized as follows. In Sect. 2 we outline the LC and HM model formulations and the MLE of the model parameters through the EM algorithm. In Sect. 3 we provide details on the proposed T-EM algorithm for both models. In Sect. 4 we summarize the main findings of an extensive simulation study aimed to assess the performance of the proposal by comparing it with the standard EM algorithm for many different scenarios. We also evaluate the proposed algorithm in connection with different initialization strategies, and compare the overall computing time. In Sect. 5 we apply the T-EM algorithm to estimate LC and HM models using a variety of data types. In Sect. 6 we provide some conclusions. Appendix A supplies more details on the settings used for the simulation studies, while Appendices B and C provide additional simulation results. Finally, Supplementary Information (SI) contains the full outcomes of every sample under each simulated scenario.

2 Model formulation

In the following, mainly borrowing from Bartolucci et al. (2013) we briefly summarize model notations and implementations of the standard MLE of the model parameters

carried out through the EM algorithm; see also Bartolucci et al. (2014) and Pandolfi et al. (2021).

2.1 Latent class model

Considering cross-sectional data and for a single individual, let $\mathbf{Y} = (Y_1, \dots, Y_r)'$ denote the vector of response variables; we assume that each variable Y_j is categorical with the same number c of categories, labeled from 0 to $c - 1$. Note that the formulation of the model may be easily adapted to the case of continuous response variables. The LC model relies on a single latent variable U with k support points that identify the latent classes in the population, labeled from 1 to k . According to the assumption of local independence, the response variables are conditionally independent given the latent variable. The model parameters are the weight of each latent class, denoted by $\pi_u = p(U = u)$, $u = 1, \dots, k$, and the conditional probability of each response variable given the latent variable, denoted by $\phi_{jy|u} = p(Y_j = y|U = u)$, for $y = 0, \dots, c - 1$, $j = 1, \dots, r$, and $u = 1, \dots, k$.

In order to estimate the model parameters, collected in the vector $\boldsymbol{\theta}$, on the basis of a sample of n independent observations \mathbf{y}_i , $i = 1, \dots, n$, the *incomplete data log-likelihood* denoted as $\ell(\boldsymbol{\theta})$ is maximized considering the *complete data log-likelihood*, given by

$$\ell^*(\boldsymbol{\theta}) = \sum_{j=1}^r \sum_{u=1}^k \sum_{y=0}^{c-1} a_{juy} \log \phi_{jy|u} + \sum_{u=1}^k b_u \log \pi_u,$$

where $a_{juy} = \sum_{i=1}^n I(u_i = u, y_{ij} = y)$ is the frequency of subjects that are in latent class u and respond by y at the j -th response variable, and $b_u = \sum_{i=1}^n I(u_i = u)$ is the number of sample units in latent class u , with $I(\cdot)$ denoting the indicator function.

2.2 Hidden Markov model

With reference to longitudinal data and for a single individual, let $\mathbf{Y}^{(t)} = (Y_1^{(t)}, \dots, Y_r^{(t)})'$ denote the occasion-specific response variables for each time $t = 1, \dots, T$, and let \mathbf{Y} denote the vector of responses, which is made of the union of the vectors $\mathbf{Y}^{(t)}$, $t = 1, \dots, T$. Given a latent process $\mathbf{U} = (U^{(1)}, \dots, U^{(T)})'$ having a discrete distribution with k states, the latent model parameters are the initial probabilities, denoted by $\pi_u = p(U^{(1)} = u)$, $u = 1, \dots, k$, and the transition probabilities denoted by $\pi_{u|\bar{u}}^{(t)} = p(U^{(t)} = u|U^{(t-1)} = \bar{u})$, $t = 2, \dots, T$, $\bar{u}, u = 1, \dots, k$. Note that it is possible to include a constraint corresponding to the hypothesis that the latent process is time homogeneous so that the transition probabilities do not depend on time occasion t : $\pi_{u|\bar{u}}^{(t)} = \pi_{u|\bar{u}}$, $t = 2, \dots, T$.

The HM model in its basic formulation (Bartolucci et al. 2013) relies on the following three main assumptions, which can be suitably relaxed:

- $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(T)}$ are conditionally independent given \mathbf{U} ;
- $Y_1^{(t)}, \dots, Y_r^{(t)}$ are conditionally independent given $U^{(t)}$, for $t = 1, \dots, T$;
- \mathbf{U} follows a first-order Markov chain with state space $\{1, \dots, k\}$, where k is the number of latent states.

2.2.1 Hidden Markov model with categorical response variables

Let $Y_j^{(t)}$, $j = 1, \dots, r$, $t = 1, \dots, T$, denote the categorical response variable with c categories, where the conditional probabilities are defined as in Section 2.1.

Given a sample of n observations, the complete data log-likelihood is expressed as

$$\ell^*(\boldsymbol{\theta}) = \sum_{j=1}^r \sum_{t=1}^T \sum_{u=1}^k \sum_{y=0}^{c-1} a_{juy}^{(t)} \log \phi_{jy|u} + \sum_{u=1}^k b_u^{(1)} \log \pi_u + \sum_{t=2}^T \sum_{\bar{u}=1}^k \sum_{u=1}^k b_{\bar{u}u}^{(t)} \log \pi_{u|\bar{u}},$$

where $a_{juy}^{(t)} = \sum_{i=1}^n I(u_i^{(t)} = u, y_{ij}^{(t)} = y)$ is the number of subjects that, at time occasion t , are in latent state u and have outcome y for the j -th response variable, $b_u^{(t)} = \sum_{i=1}^n I(u_i^{(t)} = u)$ is the number of subjects in latent state u at time occasion t , and $b_{\bar{u}u}^{(t)} = \sum_{i=1}^n I(u_i^{(t-1)} = \bar{u}, u_i^{(t)} = u)$ is the number of subjects that move from latent state \bar{u} to latent state u at time occasion t .

2.2.2 Hidden Markov model with continuous response variables

The response vectors $\mathbf{Y}^{(t)}$, $t = 1, \dots, T$, are assumed to follow a conditional Gaussian distribution, that is,

$$\mathbf{Y}^{(t)} | U^{(t)} = u \sim \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}), \quad u = 1, \dots, k,$$

with state-specific mean vectors $\boldsymbol{\mu}_u \in \mathbb{R}^r$, $u = 1, \dots, k$, and variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ constant across latent states under the assumption of homoscedasticity. This latter assumption may be relaxed to allow for heteroscedasticity across latent states.

The complete data log-likelihood function is

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

where $f(\mathbf{y}_i^{(t)} | u)$ denotes the probability density function of a multivariate Gaussian distribution with parameters $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}$, $z_{iu}^{(t)} = I(u_i^{(t)} = u)$ is an indicator function equal to 1 if subject i is in latent state u at time occasion t , and $z_{i\bar{u}u}^{(t)} = I(u_i^{(t-1)} = \bar{u}, u_i^{(t)} = u)$ is an indicator function equal to 1 if subject i moves from latent state \bar{u} to latent state u at time occasion t .

$\bar{u}, u_i^{(t)} = u$) is an indicator function equal to 1 if subject i is in latent state \bar{u} at time $t - 1$ and moves to latent state u at time t .

2.3 Expectation-maximization algorithm

Maximum likelihood estimation of model parameters is performed through the EM algorithm. Once the parameters are initialized, the EM algorithm alternates the following steps until a suitable convergence criterion is satisfied:

- **E-step:** compute the conditional expected value of $\ell^*(\theta)$ given the observed data and the value of the parameters at the previous step:

$$Q(\theta; \theta^{(h-1)}) = \mathbb{E}_{\theta^{(h-1)}}[\ell^*(\theta)|\mathbf{y}];$$

- **M-step:** maximize the expected value $Q(\theta; \theta^{(h-1)})$ and so update the model parameters:

$$\theta^{(h)} = \arg \max_{\theta} Q(\theta; \theta^{(h-1)}).$$

The computation of the expected values at the E-step is based on the following conditional probabilities, generically referred to as $q(\cdot)$. For the LC model we consider $q(u|\mathbf{y}) = p(U = u|Y = \mathbf{y})$, while for the HM model we define $q^{(t)}(u|\mathbf{y}) = p(U^{(t)} = u|Y = \mathbf{y})$ and $q^{(t)}(\bar{u}, u|\mathbf{y}) = p(U^{(t-1)} = \bar{u}, U^{(t)} = u|Y = \mathbf{y})$. In the following section, we present some details on the EM algorithm, and we show the tempering technique.

3 Tempered expectation-maximization algorithm

The T-EM algorithm is implemented by adjusting the computation of the expected frequencies in the E-step. In the following we define some general rules for the tempering constants, and we show details of the T-EM algorithm for the LC and HM models.

The family of tempered probabilities has the following expression:

$$\tilde{q}^{(\tau)}(\cdot) = m^{-1} q(\cdot)^{1/\tau}, \quad (1)$$

where $q(\cdot)$ denotes the original conditional probability, τ is a suitable parameter, known as temperature and varying over the interval $[1, +\infty)$, and m is a normalizing constant. At each E-step of the T-EM algorithm, the conditional expected frequencies are computed accordingly. Regarding the temperature, the choice $\tau \rightarrow +\infty$ yields $\tilde{q}^{(\tau)}(\cdot)$ to a uniform distribution, while $\tau = 1$ recovers the original posterior probability $q(\cdot)$. Therefore, we define a sequence of temperature values $(\tau_h)_{h \geq 1}$, where h is the algorithm iteration number, so that: (i) the initial temperature τ_1 is sufficiently large, implying that the corresponding tempered distribution $\tilde{q}^{(\tau_1)}(\cdot)$ is relatively flat and (ii)

the temperature value τ_h tends towards 1 as the algorithm iteration counter increases. The resulting sequence, denoted as *tempering profile*, guarantees a proper convergence of the algorithm (Lartigue et al. 2022).

We consider the following two tempering profiles:

- a monotonically decreasing exponential profile, which is defined as

$$\tau_h = 1 + e^{\beta - h/\alpha}, \tag{2}$$

where $\alpha \geq 1$ and $\beta \geq 0$ are two constants chosen so as to ensure flexibility in the profile shape;

- a non-monotonic profile with oscillations of gradually smaller amplitude, which is expressed as

$$\tau_h = \tanh\left(\frac{h}{2\rho}\right) + \left(\tau_0 - \beta \frac{2\sqrt{2}}{3\pi}\right) \alpha^{h/\rho} \beta \operatorname{sinc}\left(\frac{3\pi}{4} + \frac{h}{\rho}\right), \tag{3}$$

with constants $\beta, \rho, \tau_0 > 0$ and $0 < \alpha < 1$. This profile has more parameters to tune, but it guarantees a very high level of flexibility. Here $\tanh(\cdot)$ indicates the hyperbolic tangent, while $\operatorname{sinc}(x) = \sin(\pi x)/(\pi x)$ (with $\operatorname{sinc}(0) = 1$) denotes the normalized sine cardinal function. In this case, the sequence $(\tau_h)_{h \geq 1}$ may assume values that are smaller than 1 or even negative; although this is not an issue from a strictly mathematical perspective, a tempering step with negative temperature lacks a proper interpretation. Therefore, in practice, we can force the tempering profile to be always greater than or equal to 1 by taking $\tau_h = \max\{\tau_h, \delta\}$, with $\delta \geq 1$ (in this work we fix $\delta = 1$).

The abbreviations M-T-EM and O-T-EM are used for monotonic (2) and oscillating (3) tempering profiles, respectively.

3.1 Tuning of tempering profiles

The selection of optimal tempering constants for both profiles may be carried out through a grid-search procedure; in the following, the term *grid* will denote the sequence of values considered for a constant, while the term *step-size* will refer to the distance between two consecutive values.

For the monotonic profile the only two constants are simple to interpret: β controls the value of the initial temperature, while α adjusts the decrease rate of the temperature. Lower values of both make the contribution of tempering insignificant; at the extreme, $\alpha = 1$ and $\beta = 0$ recover the standard EM algorithm. Although it is not possible to provide precise and rigorous rules for the selection of these constants, some guidelines hold in general: (i) avoid very high values of α and β . Indeed, beyond certain values, the target function can not be flattened further, and only the computational time would increase. This sort of “threshold” values are unfortunately data-dependent, but we recommend not exceeding $\alpha = 15$ and $\beta = 5$; (ii) choose step-sizes for each grid such that the distance between two consecutive values of α will result much smaller

than the one between two successive values of β . Indeed, the monotonic profile is much more sensitive to variations in α than in β ; we suggest, for example, a ratio of about 1:10; (iii) avoid increasing β without a corresponding growth of α (while the opposite has no shortcomings). This would lead to a fast decrease in the value of the temperature; accordingly, the target function would not be warped back to its original shape in a gradual way, and the algorithm could possibly be brought far from the global mode; (iv) typically, for each type of data there are many possible suitable tempering configurations, and an important step is to locate a rough range for the constants. After that, although the tuning process can be further refined, most of the tempering configurations chosen within that range would provide good results; (v) various factors such as number of observations, of response variables, and of latent components would guide the choice of this “unrefined” range. For example, estimating a model with many latent components typically requires higher values of α and β with respect to a model with fewer components.

The same guidelines illustrated above should also be taken into account for the oscillating profile, where, however, there are more constants to tune. Their practical interpretation is, in this case, slightly different: T_0 controls the initial temperature, ρ the distance between two consecutive peaks of the profile, β the amplitude of the oscillations, and α the global decrease rate.

The following steps for tuning the tempering profile are derived from the aforementioned rules and are successfully employed to estimate the models for the applications presented in Sect. 5:

- (1) define grids for all the tempering constants, starting with large step-sizes;
- (2) estimate the model using the T-EM algorithm with these “unrefined” grids for the tempering constants employing a much smaller number of starting values with respect to that required with the standard EM algorithm;
- (3) identify the optimal tempering constants by comparing values of the log-likelihood function at convergence;
- (4) improve the tuning procedure, if necessary, in a smaller region of the tempering constants space and repeat the same procedure (points 2 and 3) using the same small number of different starting values.

A final note, which is effective for both profiles, is that in order to achieve a proper convergence, the algorithm needs to be run until the temperature is steadily close to 1. After that, the last step is conducted with the temperature precisely equal to 1 in order to retrieve the shape of the original log-likelihood function. Typically, this approach increases the number of steps that are required for the algorithm to converge, especially in the case of the oscillating profile. The code written for this proposal is implemented in R and it is freely available at the following link in the GitHub repository: <https://github.com/LB1304/T-EM>.

3.2 T-EM algorithm for the latent class model with categorical response variables

In the following, we provide some details of the tempered distribution (1) defined for the LC model with categorical response variables considering a suitable tempering profile τ_h :

$$\tilde{q}^{(\tau_h)}(u|y_i) = \frac{q(u|y_i)^{1/\tau_h}}{\sum_{v=1}^k q(v|y_i)^{1/\tau_h}}.$$

The corresponding pseudo-code is shown in the box Algorithm 1. The E- and M-step of the T-EM algorithm are implemented as follows:

- **E-step:** compute the conditional expected values of a_{juy} and b_u revised according to the rules

$$\tilde{b}_u^{(\tau_h)} = \sum_{i=1}^n \tilde{q}^{(\tau_h)}(u|y_i) \quad \text{and} \quad \tilde{a}_{juy}^{(\tau_h)} = \sum_{i=1}^n I(y_{ij} = y) \tilde{q}^{(\tau_h)}(u|y_i)$$

to obtain the conditional expected value $Q(\theta; \theta^{(h-1)})$.

- **M-step:** maximize $Q(\theta; \theta^{(h-1)})$, thus updating the parameters as:

$$\pi_u^{(\tau_h)} = \frac{\tilde{b}_u^{(\tau_h)}}{n} \quad \text{and} \quad \phi_{jy|u}^{(\tau_h)} = \frac{\tilde{a}_{juy}^{(\tau_h)}}{\tilde{b}_u^{(\tau_h)}}.$$

Algorithm 1 T-EM algorithm for LC model with categorical response variables

- 1: Define a tempering profile $(\tau_h)_{h \geq 1}$.
 - 2: $\theta \leftarrow \theta^{(0)}$ and $h \leftarrow 0$.
 - 3: **while** (Convergence Condition = FALSE) **do**
 - 4: $h \leftarrow h + 1$;
 - 5: **E-Step:** compute $\tilde{a}_{juy}^{(\tau_h)}$ and $\tilde{b}_u^{(\tau_h)}$;
 - 6: **M-Step:** compute $\pi_u^{(\tau_h)}$ and $\phi_{jy|u}^{(\tau_h)}$.
 - 7: **end while**
-

3.3 T-EM algorithm for the hidden Markov model with categorical response variables

A more refined formulation for the tempered distribution in (1) is required to estimate the HM model. Once the tempering profile τ_h is chosen, we obtain the following tempered distributions:

$$\tilde{q}^{(t; \tau_h)}(u|y_i) = \frac{q^{(t)}(u|y_i)^{1/\tau_h}}{\sum_{v=1}^k q^{(t)}(v|y_i)^{1/\tau_h}}$$

and

$$\tilde{q}^{(t; \tau_h)}(\bar{u}, u | \mathbf{y}_i) = \frac{q^{(t)}(\bar{u}, u | \mathbf{y}_i)^{1/\tau_h}}{\sum_{\bar{v}=1}^k \sum_{v=1}^k q^{(t)}(\bar{v}, v | \mathbf{y}_i)^{1/\tau_h}}.$$

The pseudo-code is shown in the box Algorithm 2. In this setting, the steps of the T-EM algorithm are:

- **E-step:** compute the revised conditional expected value of every frequency $a_{juy}^{(t)}$, $b_u^{(t)}$, and $b_{\bar{u}u}^{(t)}$, so as to obtain the conditional expected value $\mathcal{Q}(\theta; \theta^{(h-1)})$; in particular, we have the following explicit expressions:

$$\begin{aligned} \tilde{a}_{juy}^{(t; \tau_h)} &= \sum_{i=1}^n I(y_{ij}^{(t)} = y) \tilde{q}^{(t; \tau_h)}(u | \mathbf{y}_i), \\ \tilde{b}_u^{(t; \tau_h)} &= \sum_{i=1}^n \tilde{q}^{(t; \tau_h)}(u | \mathbf{y}_i), \\ \tilde{b}_{\bar{u}u}^{(t; \tau_h)} &= \sum_{i=1}^n \tilde{q}^{(t; \tau_h)}(\bar{u}, u | \mathbf{y}_i). \end{aligned}$$

Similarly to the standard EM algorithm, posterior probabilities $\tilde{q}^{(t; \tau_h)}(u | \mathbf{y}_i)$ and $\tilde{q}^{(t; \tau_h)}(\bar{u}, u | \mathbf{y}_i)$ may be efficiently computed by a backward recursion; see Bartolucci et al. (2013, pp 61–64) for further details.

- **M-step:** by maximizing $\mathcal{Q}(\theta; \theta^{(h-1)})$ update the parameters as follows:

$$\pi_u^{(\tau_h)} = \frac{\tilde{b}_u^{(1; \tau_h)}}{n}, \quad \pi_{u|\bar{u}}^{(t; \tau_h)} = \frac{\tilde{b}_{\bar{u}u}^{(t; \tau_h)}}{\tilde{b}_{\bar{u}}^{(t-1; \tau_h)}}, \quad \text{and} \quad \phi_{jy|u}^{(\tau_h)} = \frac{\sum_{t=1}^T \tilde{a}_{juy}^{(t; \tau_h)}}{\sum_{t=1}^T \tilde{b}_u^{(t; \tau_h)}}.$$

Algorithm 2 T-EM algorithm for HM model with categorical response variables

- 1: Define a tempering profile $(\tau_h)_{h \geq 1}$.
 - 2: $\theta \leftarrow \theta^{(0)}$ and $h \leftarrow 0$.
 - 3: **while** (Convergence Condition = FALSE) **do**
 - 4: $h \leftarrow h + 1$;
 - 5: **E-Step:** compute $\tilde{a}_{juy}^{(t; \tau_h)}$, $\tilde{b}_u^{(t; \tau_h)}$, and $\tilde{b}_{\bar{u}u}^{(t; \tau_h)}$;
 - 6: **M-Step:** compute $\pi_u^{(\tau_h)}$, $\pi_{u|\bar{u}}^{(t; \tau_h)}$, and $\phi_{jy|u}^{(\tau_h)}$.
 - 7: **end while**
-

3.4 T-EM algorithm for hidden Markov model with continuous response variables

Regarding the HM model with continuous response variables, the pseudo-code is shown in the box Algorithm 3. Similarly to the previous case, the steps of the resulting T-EM algorithm are as follows:

- **E-step:** compute the conditional expected value $\mathcal{Q}(\theta; \theta^{(h-1)})$ considering $z_{iu}^{(t)}$ and $z_{i\bar{u}\bar{u}}^{(t)}$:

$$\tilde{z}_{iu}^{(t; \tau_h)} = \tilde{q}^{(t; \tau_h)}(u | y_i) \quad \text{and} \quad \tilde{z}_{i\bar{u}\bar{u}}^{(t; \tau_h)} = \tilde{q}^{(t; \tau_h)}(\bar{u}, u | y_i).$$

- **M-step:** maximize $\mathcal{Q}(\theta; \theta^{(h-1)})$ and update the model parameters as follows:

$$\begin{aligned} \mu_u^{(\tau_h)} &= \frac{1}{\sum_{i=1}^n \sum_{t=1}^T \tilde{z}_{iu}^{(t; \tau_h)}} \sum_{i=1}^n \sum_{t=1}^T \tilde{z}_{iu}^{(t; \tau_h)} y_i^{(t)}, \\ \Sigma^{(\tau_h)} &= \frac{\sum_{i=1}^n \sum_{t=1}^T \sum_{u=1}^k \tilde{z}_{iu}^{(t; \tau_h)} (y_i^{(t)} - \mu_u)(y_i^{(t)} - \mu_u)'}{nT}, \\ \pi_u^{(\tau_h)} &= \frac{\sum_{i=1}^n \tilde{z}_{iu}^{(1; \tau_h)}}{n}, \\ \pi_{u|\bar{u}}^{(t; \tau_h)} &= \frac{\sum_{i=1}^n \sum_{t=2}^T \tilde{z}_{i\bar{u}\bar{u}}^{(t; \tau_h)}}{\sum_{i=1}^n \sum_{t=2}^T \tilde{z}_{iu}^{(t-1; \tau_h)}}. \end{aligned}$$

Algorithm 3 T-EM algorithm for HM model with continuous response variables

- 1: Define a tempering profile $(\tau_h)_{h \geq 1}$.
 - 2: $\theta \leftarrow \theta^{(0)}$ and $h \leftarrow 0$.
 - 3: **while** (Convergence Condition = FALSE) **do**
 - 4: $h \leftarrow h + 1$;
 - 5: **E-Step:** compute $\tilde{z}_{iu}^{(t; \tau_h)}$ and $\tilde{z}_{i\bar{u}\bar{u}}^{(t; \tau_h)}$;
 - 6: **M-Step:** compute $\mu_u^{(\tau_h)}$, $\Sigma^{(\tau_h)}$, $\pi_u^{(\tau_h)}$, and $\pi_{u|\bar{u}}^{(t; \tau_h)}$.
 - 7: **end while**
-

4 Simulation study

We conducted an extensive Monte Carlo simulation study to evaluate the performance of the T-EM algorithm. In the following, we illustrate the simulation schemes for each different model specifications and summarize the main results.

4.1 Settings of the experimental scenarios

The settings involved in each model are different values of sample size n , number of response variables r , categories for each variable c , time occasions T , and latent components k . We define a baseline scenario (setting A, see Tables 16, 17, and 18 in Appendix A) for each model, characterized by $n = 500$, $r = 6$, $c = 3$, $T = 5$, and $k = 3$. In addition, more scenarios (settings from B to F in Appendix A) are

obtained by doubling, one at a time, the value of each feature. In Tables 16, 17, and 18 in Appendix A also the values of the models' parameters are presented. For each scenario, 50 different samples are drawn. For each of the simulated samples, we estimate 100 times both the model with correctly specified latent structure and that with misspecified latent structure, using each time different starting values randomly selected and employing the standard EM algorithm and the two proposed versions of the T-EM algorithm. The choice to also fit misspecified models allows us to show in more detail the features of the proposed tempering approach.

The convergence of the algorithms is checked on the basis of both the relative change in the log-likelihood of two consecutive steps, and the distance between the corresponding parameter vectors. We stop the algorithm when both criteria are satisfied:

$$\frac{\ell(\boldsymbol{\theta}^{(h)}) - \ell(\boldsymbol{\theta}^{(h-1)})}{|\ell(\boldsymbol{\theta}^{(h)})|} < \varepsilon_1$$

and

$$\max_s |\theta_s^{(h)} - \theta_s^{(h-1)}| < \varepsilon_2,$$

where $\boldsymbol{\theta}^{(h)}$ is the vector of parameter estimates obtained at the h -th iteration of the M-step and ε_1 and ε_2 are tolerance levels equal to 10^{-8} and 10^{-4} , respectively.

Regarding the algorithm initialization, we adopt a starting rule based on normalized random numbers (Bartolucci et al. 2013). In more details, each initial (π_u) and transition ($\pi_{u|\bar{u}}^{(t)}$) probability is initialized with a random number drawn from a uniform distribution between 0 and 1. Then, they are normalized so that $\sum_{u=1}^k \pi_u = 1$ and $\sum_{u=1}^k \pi_{u|\bar{u}}^{(t)} = 1$. Similarly, we draw each $\phi_{jy|u}$ from the uniform distribution and we normalize these parameters so that $\sum_{y=0}^{c-1} \phi_{jy|u} = 1$. In the case of continuous response variables, the mean vectors $\boldsymbol{\mu}_u$ are drawn from a multivariate Gaussian distribution, whereas $\boldsymbol{\Sigma}$ is initialized with the observed variance-covariance matrix. As suggested in Bartolucci et al. (2013), combining deterministic and random starting values is a proper approach. Therefore, in Sect. 4.4 we analyze the behavior of tempering in connection with a different initialization strategy.

4.2 Simulation results

The EM and T-EM algorithms are compared according to the following criteria:

1. *Global maximum achievement*: the highest of the maximized log-likelihood values over all 100 initial values, denoted by $\hat{\ell}_{\text{MAX}}$, is considered as the global maximum, and a log-likelihood value at convergence denoted by $\hat{\ell}$ is considered close to this value once it satisfies $(\hat{\ell}_{\text{MAX}} - \hat{\ell})/|\hat{\ell}_{\text{MAX}}| < \tilde{\varepsilon}$, where $\tilde{\varepsilon}$ is a suitable threshold;
2. *Average distance from the global maximum* computed over the 100 log-likelihood values $\hat{\ell}_1, \dots, \hat{\ell}_{100}$ and expressed as $\sum_{s=1}^{100} (\hat{\ell}_{\text{MAX}} - \hat{\ell}_s)/100$;

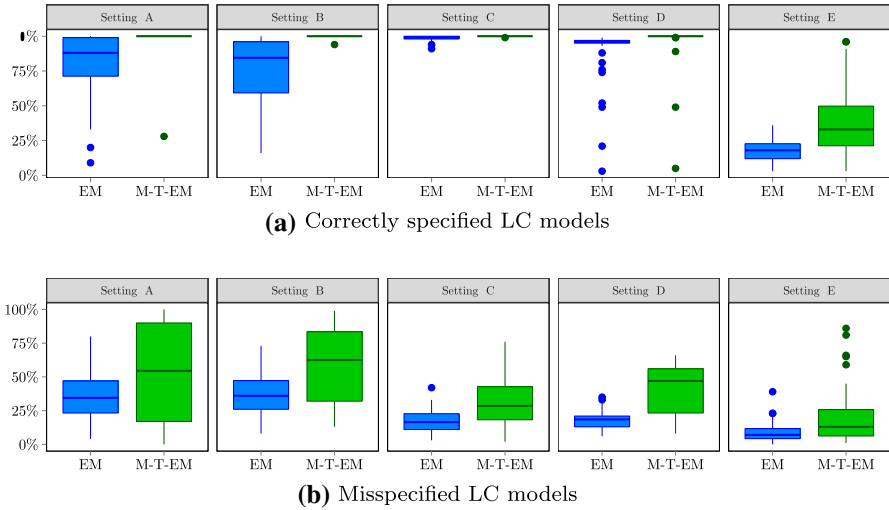


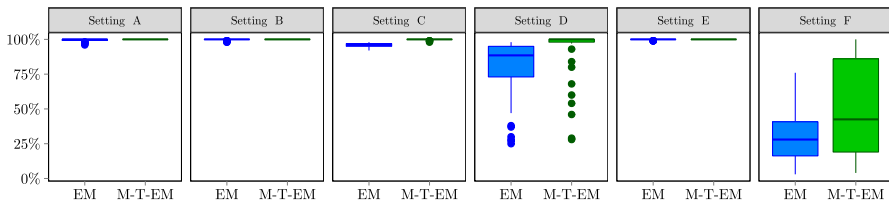
Fig. 1 Percentages of global maxima obtained using EM and M-T-EM algorithms under the simulated scenarios presented in Table 16 of the Appendix A for the LC model

3. *Low mean square error* of the estimated model parameters with respect to the true model parameters, computed only for models with a correctly specified latent structure;
4. *Low mean and median of the log-likelihood values at convergence.*

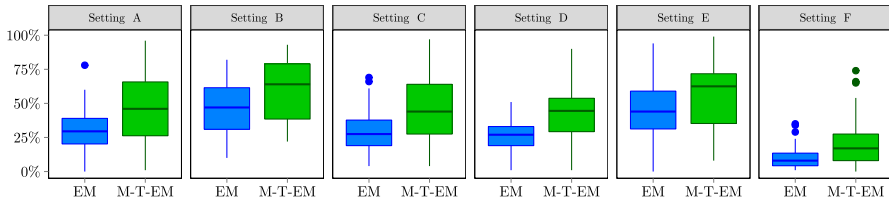
In particular, in this first part of the simulation study, we analyze the performance of the M-T-EM algorithm when the tempering profile is optimally tuned through a grid-search procedure. The following values for the tempering constants are kept fixed throughout the simulation studies: α ranging from 1 to 15 with a step-size equal to 1 and β ranging from 0 to 2, with a step-size equal to 0.1. In order to show the flexibility of the method, we use the same grid for each model. However, efficient ad hoc grids may be set according to the model and observed data. The results are summarized in the following, and the full outcomes related to every sample under each simulated scenario are reported in the SI.

Criterion 1 is the most important, providing a suitable measure of performance of the algorithm. In this regard, the main results are summarized in Figs. 1, 2, and 3, representing the frequencies of global maximum with respect to the LC model, HM model with categorical response variables, and HM model with continuous response variables, respectively. From all these figures it clearly emerges that the M-T-EM algorithm ensures better performance in each considered scenario.

Regarding the estimation of models whose latent structure is correctly specified, in particular (see Figs. 1a, 2a, and 3a), the improvement with respect to the standard EM algorithm is very relevant: the M-T-EM is generally able to detect the global maximum in the overwhelming majority of cases, and the frequency of convergence to the global mode is very close, or even equal, to 100%. Only in estimating models with many latent states (up to 6), this percentage is slightly reduced, even if the M-T-EM still remains the algorithm providing the best performance. As an example, we consider the

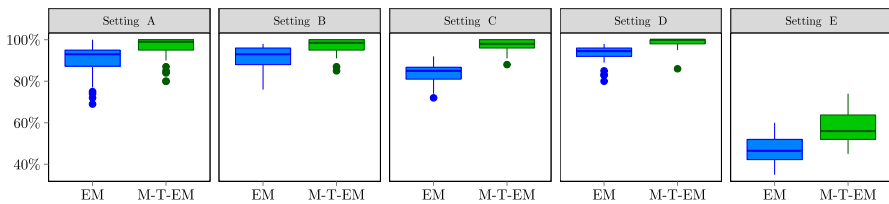


(a) Correctly specified HM models with categorical response variables

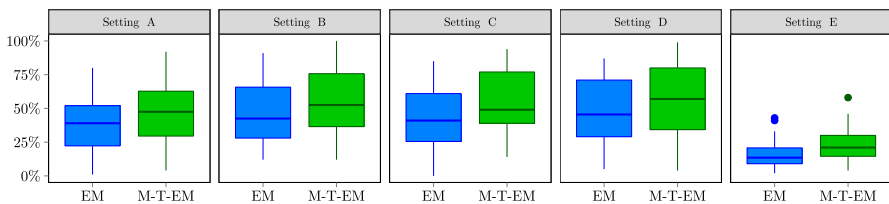


(b) Misspecified HM models with categorical response variables

Fig. 2 Percentages of global maximum using EM and M-T-EM algorithms under the simulated scenarios presented in Table 17 of the Appendix A for the HM model with categorical response variables



(a) Correctly specified HM models with continuous response variables



(b) Misspecified HM models with continuous response variables

Fig. 3 Percentages of global maximum using EM and M-T-EM algorithms under the simulated scenarios presented in Table 18 of the Appendix A for the HM model with continuous response variables

HM model with categorical response variables and in the particular setting F (see the last plot in Fig. 2a): in this case the frequency of convergence to the global maximum is, on average, equal to 29% when the standard EM algorithm is used, and up to 52% when the M-T-EM algorithm is employed. Moreover, this frequency is always lower than 75% with the EM, while it reaches 100% with M-T-EM (though only in a few cases).

Table 1 Number of samples in which the global maximum is reached with frequency < 10%, > 50%, or > 95%, using EM (highlighted in bold) and M-T-EM (highlighted in italic) algorithms under the simulated scenarios presented in Table 16 of the Appendix A for the LC model

	Correctly specified			Misspecified		
	< 10%	> 50%	> 95%	< 10%	> 50%	> 95%
A	<i>1-0</i>	43-49	20-49	7-2	9-25	0-8
B	0-0	41-50	15-49	1-0	10-33	0-3
C	0-0	50-50	47-50	9-4	0-9	0-0
D	<i>1-1</i>	47-48	32-47	4-1	0-21	0-0
E	10-5	0-13	0-2	32-18	0-5	0-0

Table 2 Number of samples in which the global maximum is reached with frequency < 10%, > 50%, or > 95%, using EM (highlighted in bold) and M-T-EM (highlighted in italic) algorithms under the simulated scenarios presented in Table 17 of the Appendix A for the HM model with categorical response variables

	Correctly specified			Misspecified		
	< 10%	> 50%	> 95%	< 10%	> 50%	> 95%
A	0-0	50-50	50-50	4-1	6-24	0-1
B	0-0	50-50	50-50	0-0	20-35	0-0
C	0-0	50-50	35-50	4-1	5-21	0-1
D	0-0	43-47	11-41	3-2	1-17	0-0
E	0-0	50-50	50-50	3-1	20-32	0-2
F	7-6	6-23	0-7	27-17	0-5	0-0

Table 3 Number of samples in which the global maximum is reached with frequency < 10%, > 50%, or > 95%, using EM (highlighted in bold) and M-T-EM (highlighted in italic) algorithms under the simulated scenarios presented in Table 18 of the Appendix A for the HM model with continuous response variables

	Correctly specified			Misspecified		
	< 10%	> 50%	> 95%	< 10%	> 50%	> 95%
A	0-0	50-50	12-36	3-2	13-23	0-0
B	0-0	50-50	14-36	0-0	20-26	0-1
C	0-0	50-50	0-40	2-0	19-24	0-0
D	0-0	50-50	18-47	4-4	20-28	0-1
E	0-0	15-40	0-0	17-5	0-1	0-0

All the algorithms are less efficient in steadily detecting the global mode when models with misspecified latent components are estimated (see Figs. 1b, 2b, and 3b). The M-T-EM algorithm always provides the best performance, and in many scenarios the improvement is very relevant: in setting D of the LC model (Fig. 1b) the frequency of convergence to the global mode increases from 18 to 41%; in setting C of the HM model with categorical responses (Fig. 2b) for some samples this frequency reaches 100%.

In Tables 1, 2, and 3, for each one of the simulated scenarios, we show the number of samples in which the global maximum is reached at least half of the times (> 50%), almost always (> 95%), or almost never (< 10%). These results provide supporting evidence for the conclusions drawn so far. In particular, when the considered models are estimated with the correct latent structure, the M-T-EM algorithm performs really well, and significantly better than the standard EM algorithm. For example, this enhancement is evident in setting C of the HM model with continuous response vari-

Table 4 Mean square errors of the estimated model parameters with respect to the true model parameters, using EM (highlighted in bold) and M-T-EM (highlighted in italic) algorithms under simulated scenarios presented in Tables 16, 17, and 18 in the Appendix A and estimating models with correct latent structure

Scenario	LC	Categorical HM	Continuous HM
A	0.0013 – <i>0.0012</i>	0.0006 – <i>0.0002</i>	0.0643 – <i>0.0272</i>
B	0.0007 – <i>0.0006</i>	0.0003 – <i>0.0001</i>	0.0556 – <i>0.0294</i>
C	0.0022 – <i>0.0010</i>	0.0046 – <i>0.0003</i>	0.1603 – <i>0.0433</i>
D	0.0020 – <i>0.0006</i>	0.0027 – <i>0.0002</i>	0.0322 – <i>0.0094</i>
E	0.0584 – <i>0.0544</i>	0.0002 – <i>0.0001</i>	0.1384 – <i>0.1168</i>
F	–	0.0202 – <i>0.0179</i>	–

ables, where we observe that 40 samples reach the global mode with high frequency compared to none with the standard EM algorithm. An analogous improvement is noticeable for the case with 6 latent states but referred to the frequency of convergence to the global maximum more than half the time. In the case of models estimated with the wrong latent structure and many components, we show another important result, not highlighted so far: the number of samples in which the global maximum is almost never reached ($< 10\%$ of times) diminishes when the M-T-EM algorithm is employed.

We also consider the mean distance from the global mode to measure how far the obtained maximum is from the global one. In particular, although all settings provide similar results, we notice that when dealing with correctly specified models, the mean distance decreases to zero when the M-T-EM algorithm is employed, thus confirming that the global maximum is almost always reached. In Appendix B, all detailed results are provided in Figs. 7, 8, and 9.

Finally, we also provide the mean square error of the estimated model parameters with respect to the true values, once the models are estimated with the correct latent structure. The results, summarized in Table 4, show that the mean square error values are always smaller with the M-T-EM algorithm than with the standard EM algorithm, thus highlighting that the estimated model parameters are more accurate by employing the former.

4.3 Results in terms of computational time

Having assessed the good performance of the proposed M-T-EM algorithm in locating the global maximum, we also compare the computational time required for convergence with that required by the EM algorithm for the same simulation settings illustrated above. Tempering constants are chosen as presented in Sect. 4.2. The estimation is performed by employing an Intel(R) Core(TM) i7-8700T CPU @ 2.40GHz Windows desktop with 8 GB of RAM.

The main results, summarized in Table 5, show that when estimating LC and HM models with continuous response variables, the EM and M-T-EM algorithms show very similar computing times. The EM algorithm generally remains the fastest even if

Table 5 Computational time in seconds of the EM (highlighted in bold) and M-T-EM (highlighted in italic) algorithms for each settings, computed as the mean over 50 samples and 100 starting values as presented in Sect. 4.2

Scenario	LC	Categorical HM	Continuous HM
Correctly specified models			
A	0.039 –0.055	0.178 –0.643	3.499 –3.442
B	0.042 –0.067	0.288 –1.109	6.225 –6.381
C	0.046 –0.052	0.212 –0.583	7.475 –7.224
D	0.035 –0.039	0.191 –0.775	5.974 –5.897
E	0.466 –0.537	0.270 –1.206	9.545 –9.495
F	–	1.728 –11.237	–
Misspecified models			
A	0.205 –0.484	1.114 –7.282	11.114 –12.480
B	0.268 –0.348	2.045 –13.258	18.646 –21.016
C	0.294 –0.407	1.396 –6.747	24.670 –29.081
D	0.244 –0.364	1.173 –7.365	19.981 –23.852
E	0.581 –0.630	2.022 –13.217	18.513 –19.714
F	–	2.523 –15.867	–

















the difference with the M-T-EM is negligible. When dealing with correctly specified HM models with continuous response variables, the M-T-EM algorithm is faster than the EM algorithm. Conversely, for the case of the HM model with categorical response variables it is the slowest, requiring up to 6.5 times the computational time of the EM algorithm. These two opposite behaviors are due to the different implementations of the T-EM algorithm: the one for the HM model with categorical responses requires an additional loop to the code with respect to the other two models.

4.4 Initialization of the T-EM algorithm

In this section we consider different initialization strategies of the model parameters to evaluate the effect of the different choices in detecting the global maximum and reducing the computational time. For continuous data, as proposed by Leroux and Puterman (1992), and following McLachlan and Basford (1988), we initialize the parameters according to the partition obtained by applying the k -means method (MacQueen 1967). Maruotti and Punzo (2021) inspected this initialization approach and a few others, concluding that the k -means strategy provides the best results. A similar initialization is employed for discrete data applying the k -modes algorithm (Huang 1998). Initial values are computed as follows:

- proportion of observations assigned to cluster u at the first time occasion for the initial probabilities (π_u);
- proportion of transition (or persistence) estimated from cluster \bar{u} to cluster u for the transition probabilities ($\pi_{u|\bar{u}}^{(t)}$);

Table 6 Percentage of samples in which the global maximum is reached by the M-T-EM algorithm with k -means initialization, but not by the standard EM algorithm with the same starting values when the latent structure of the models is correctly specified

	LC	Categorical HM	Continuous HM
A	98% 	62% 	14% 
B	98% 	58% 	8% 
C	96% 	0% 	0% 
D	76% 	76% 	0% 
E	60% 	38% 	28% 
F	–	84% 	–

- proportion of observations assigned to cluster u who responded with category y to the response variable j for the conditional probabilities ($\phi_{jy|u}$);
- maximum likelihood estimator on the observations of cluster u for the mean vectors (μ_u);
- maximum likelihood estimator on all the observations under the hypothesis of homoscedasticity for the variance-covariance matrix (Σ).

We consider the same samples and starting values used in Sect. 4.2, comparing the performance of the EM and the M-T-EM algorithms. In general, when the estimation of correctly specified models is considered, the standard EM algorithm benefits from the adoption of a k -means initialization using this kind of strategy, therefore, the results obtained with the EM and the M-T-EM algorithms are very similar.

In Table 6, for each scenario, we report the number of samples in which the standard EM algorithm with k -means initialization does not converge to the global maximum, which is instead reached by the M-T-EM algorithm with the same starting values. It is important to remark that M-T-EM algorithm does not behave worse than the standard EM algorithm in all the other samples, but both algorithms converge to the same value. Further analyses conducted on correctly specified HM models with continuous response variables and $k = 2$ latent states highlight that in such case the global maximum is always reached also by the EM algorithm with k -means initialization.

We also compare random and k -means initializations for the M-T-EM algorithm. The results, summarized in Table 7, show that the k -means initialization works properly. Indeed this strategy significantly reduces the number of iterations required for convergence, and hence the computational time. In particular we report, along with the number of samples in which the M-T-EM algorithm with k -means initialization reaches the global maximum, the average number of iterations required by the two initialization strategies to converge. We notice that apart from some cases with many latent components, the global maximum is almost always reached by the M-T-EM algorithm when initialized with the k -means approach. As for the decrease in the number of iterations, the advantage is particularly evident when dealing with HM model with continuous responses; in this case, it is dropped up to one sixth.

In the case of models where the latent structure is not correctly specified, the situation is less well defined: likewise the previous case, the results obtained comparing EM and M-T-EM algorithms initialized with k -means strategy are very similar for some samples (Table 8), highlighting that the standard EM algorithm may sometimes benefit

Table 7 Percentage of samples in which the M-T-EM algorithm with k -means initialization reaches the global maximum and number of iterations until convergence with random and k -means (or k -modes) initialization when the latent structure of the models is correctly specified

















	Percentage (Glob. Max)		Iterations (Random)	Iterations (k -means)
LC model				
A	98%		26.49	25.10
B	98%		28.88	28.32
C	100%		11.00	6.82
D	92%		11.42	8.54
E	0%		–	–
HM model (categorical responses)				
A	100%		10.09	5.48
B	100%		10.00	5.14
C	100%		8.89	5.78
D	92%		12.53	9.70
E	100%		9.50	5.16
F	36%		176.47	164.96
HM model (continuous responses)				
A	100%		42.73	11.84
B	100%		37.97	10.76
C	100%		45.39	11.06
D	98%		34.21	10.62
E	100%		83.89	14.44

Table 8 Percentage of samples in which the global maximum is reached by the M-T-EM algorithm with k -means initialization, but not by the standard EM algorithm with the same starting values when the latent structure of the models is not correctly specified

































Scenario	LC	Categorical HM	Continuous HM
A	60% 	86% 	44% 
B	60% 	76% 	70% 
C	38% 	82% 	62% 
D	60% 	42% 	68% 
E	78% 	70% 	62% 
F	–	78% 	–

Table 9 Percentage of samples in which the M-T-EM algorithm with k -means initialization reaches the global maximum and number of iterations until convergence with random and k -means (or k -modes) initialization when the latent structure of the models is not correctly specified

Scenario	Percentage (Glob. Max)		Iterations (Random)	Iterations (k -means)
LC model				
A	16%		359.44	216.00
B	22%		136.21	121.27
C	0%		–	–
D	26%		112.04	99.04
E	0%		–	–
HM model (categorical responses)				
A	40%		148.55	140.29
B	40%		134.40	121.61
C	32%		122.13	110.48
D	30%		147.47	132.06
E	32%		116.50	106.26
F	18%		263.00	253.79
HM model (continuous responses)				
A	44%		142.33	142.33
B	56%		117.07	97.32
C	50%		141.78	136.68
D	44%		116.41	94.09
E	26%		136.21	89.54

from the adoption of this initialization strategy. However, when M-T-EM is employed, this improvement does not always correspond to an advantage when using the k -means initialization with respect to the random one. As shown in Table 9, the number of samples that benefit from this initialization strategy is quite limited and usually does not reach the 50%. Finally, also in this case the k -means initialization provides some benefits from the point of view of the number of iterations until convergence, even if less pronounced than in the case of models with correctly specified latent structures.

4.5 The role of the oscillating tempering profile

Although the M-T-EM algorithm ensures significant improvements in terms of ability to detect the global maximum, in some cases the frequency of convergence to this global mode remains inferior to 100%. A possible remedy is represented by the oscillating profile, which is able to explore the parameter space more deeply than the monotonic one. In the following we focus only on the LC model, comparing the O-T-EM algorithm with the EM and M-T-EM algorithms; this is due to the higher computing time associated with this profile. The main results are summarized in Fig. 4,

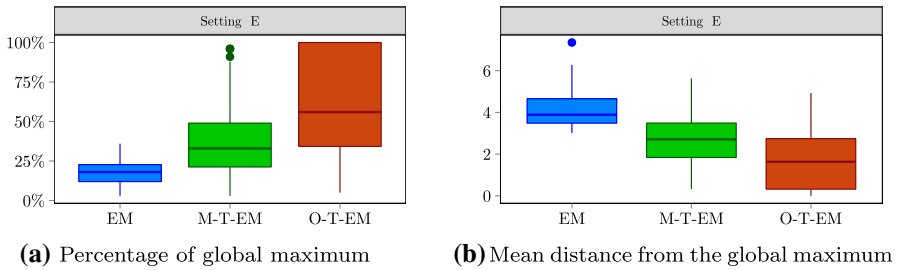


Fig. 4 Percentage of global maximum and mean distance from it with the EM, M-T-EM, and O-T-EM algorithms on simulated data from an LC model correctly specified with six latent classes

Table 10 Computational time in seconds of the EM, M-T-EM, and O-T-EM algorithms, computed as the mean over 50 samples and 100 starting values, as presented in Sect. 4.2

Algorithm	Minimum	Median	Mean	Maximum
EM	0.07	0.51	0.466	1.74
M-T-EM	0.11	0.59	0.537	1.78
O-T-EM	0.08	6.08	7.91	24.51

where we show the percentage of times the global maximum is reached and the mean distance from the global maximum for the three versions of the algorithm.

Employing the oscillating profile, we notice a further improvement compared to the results analyzed in Sect. 4.2: the global maximum is reached on average about 18% of times with the standard EM algorithm, which increases up to 38% with the M-T-EM algorithm, and up to 60% with the oscillating version. It is also interesting to evaluate the number of samples in which the global maximum is reached almost surely ($< 95\%$); this number, as reported in Table 1, was equal to 0 and 2 with EM and M-T-EM algorithms, respectively. Using the O-T-EM algorithm instead it increases to 18 samples. As for the mean distance from the global maximum, we notice that this value decreases accordingly, following the general advantage of the O-T-EM algorithm over the monotonic version. This optimal behavior of the tempered algorithm with oscillating profile results, however, in a much higher computational time, as reported in Table 10. This aspect sometimes makes the employment of the O-T-EM algorithm rather complex; in particular, when it is applied to the HM model with categorical responses, the convergence is extremely slow, and the M-T-EM could be the most appropriate choice.

4.6 Analysis of the T-EM algorithm with fixed tempering profile

Lastly, we check the performance of the T-EM algorithm when it is not optimally tuned, but the tempering constants are fixed in advance. With this aim, for each inspected scenario, a short list of different configurations of tempering constants is considered for applying the M-T-EM algorithm to all samples. In the analysis of the results, the tempered version is considered as the best choice only when it outperforms the standard EM algorithm with respect to all the four criteria introduced in

Sect. 4.2. Otherwise, if at least one criterion shows a better result with the standard EM algorithm, the latter is preferred. In this way, we carry out a very rigorous analysis.

Tables 19, 20, and 21 in the Appendix C report for each scenario the configuration of tempering constants which exhibits the best performance. Results are highly satisfactory in most cases: given a fixed configuration, the M-T-EM algorithm outmatches the standard version in around 50% of samples in almost all the analyzed scenarios. In other words, once a configuration of tempering constants is set appropriately by a grid-search procedure over a specific sample, it generally remains valid for around 50% of other samples. This percentage increases up to 100% in some scenarios, especially when the latent structure of the model is correctly specified: the considered configuration of tempering constants provides optimal results in all samples. Similar results are achieved in the case of oscillating tempering profile analyzing setting E of the LC model when the latent structure is correctly defined: the best configuration of tempering constants ($\alpha = 0.9$, $\beta = 50$, $\rho = 5$, and $T_0 = 10$) performs well with 62% of the considered samples. It is clear that there are still some cases that require experimenting with the tempering constants to yield good performance; however, in our opinion, this represents a first significant improvement that allows avoiding specific settings for models and types of data.

5 Applications

To explore the performance of the T-EM algorithm when dealing with real-world cases, we apply it to cross-sectional and longitudinal data; we specifically address the problem of selecting the best number of components for LC and HM models.

5.1 Evaluation of anxiety and depression

We consider data derived from the administration of 14 ordinal items measuring anxiety and depression in a sample of 201 Italian oncological patients (Zigmond and Snaith 1983). Items are measured according to four response categories ranging from 0 to 3 and corresponding to the lowest and to the highest level of anxiety or depression, respectively. Data are available in the R package `MULTILCIRT` (Bartolucci et al. 2014).

The LC model allows to discover subpopulations of patients with similar intensity levels of these two pathologies. The model is estimated with both EM and T-EM algorithms with a number of latent components k ranging from 1 to 4 to perform model selection. The Bayesian Information Criterion (Schwarz 1978, *BIC*) is employed at this purpose penalizing the maximized log-likelihood function for the model complexity.

For the M- and O-T-EM algorithms the following two configurations of tempering constants are used and held fixed over the values of k : $\alpha = 42$ and $\beta = 1.5$ for the monotonic version, and $\rho = 90$, $\tau_0 = 10$, $\beta = 20$, and $\alpha = 0.8$ for the oscillating one. In the following, we show the results only for values of k for which there is a significant difference on the global maximum reached by employing the EM and the T-EM algorithms. Figure 5 refers to the maximum log-likelihood values reached by

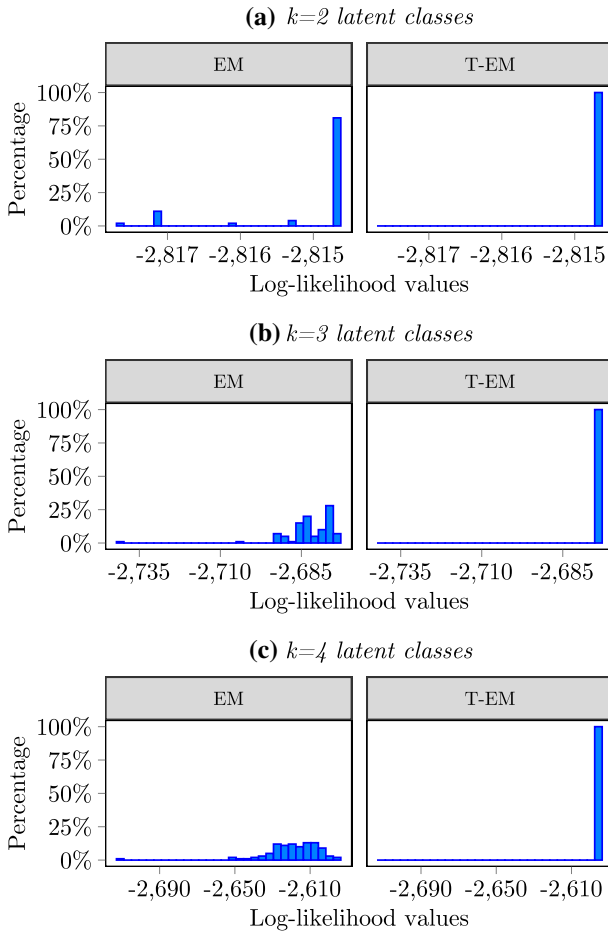


Fig. 5 Maximized log-likelihood values of the LC model for the anxiety and depression data using standard EM (left) and T-EM (right) algorithms; as for the latter, monotonic and oscillating versions provide the same results. Three different choices for the number of latent classes are analyzed, with 100 random starting values each

each algorithm for every model. As it is evident, while the EM algorithm spreads out over a wide range of values, both tempered algorithms always converge to a single value appearing as the global mode.

Results based on the O-T-EM algorithm are reported in Table 11, where it can be seen that the optimal number of components corresponding to the minimum value of BIC is three. It is important to remark that the results are always obtained using the same configuration of tempering constants as presented above. Therefore, we highlight again the considerable level of flexibility of the proposed method.

Table 11 Maximum log-likelihood, number of parameters and BIC value resulting from fitting a LC model with the O-T-EM algorithm for different values of k . The value in bold represents the best result

k	$\hat{\ell}$	#par	BIC
1	-3,153.15	42	6,529.04
2	-2,814.64	85	6,080.05
3	-2,674.48	128	6,027.79
4	-2,595.47	171	6,097.83

Table 12 Maximum log-likelihood, number of parameters and BIC index resulting from fitting a time heterogeneous HM model with the M-T-EM algorithm for different numbers of latent states k . The value in bold represents the best result

k	$\hat{\ell}$	#par	BIC
1	-27,936.35	10	55,964.81
2	-22,638.39	31	45,562.30
3	-22,275.05	62	45,121.14
4	-22,051.55	103	45,051.77
5	-21,881.36	154	45,181.12

Table 13 Mean and median of maximized log-likelihood values of the HM model, proportion (*Perc.*) of global maximum and mean distance (*Dist.*) from the global mode, using EM and M-T-EM algorithms on criminal data with $k = 4$ latent states

	Mean	Median	Perc.% [Dist.] (Glob. Max)
EM	-22,075.02	-22,051.60	73% [23.52]
M-T-EM	-22,053.53	-22,051.51	98% [2.03]

5.2 Discovering criminal trajectories

We consider longitudinal data on conviction histories of a cohort of $n = 10,000$ offenders followed from the age of criminal responsibility (10 years) until age 40. As described in Research Development and Statistics Directorate (1998), offenses are grouped into the following 10 typologies: violence against the person, sexual offenses, burglary, robbery, theft and handling stolen goods, fraud and forgery, criminal damage, drug offenses, motoring offenses, and other offenses. Binary response variables ($r = 10$) indicate if the offender has committed a crime during six age bands ($T = 6$) of length equal to five years. An HM model was proposed for the analysis of these data in Bartolucci et al. (2007) and Pennoni (2014) to identify typologies of criminal behavior and types of criminal career specialization over time.

Results of estimating a time heterogeneous HM model with the M-T-EM algorithm for a number of states ranging from 1 to 5 are reported in Table 12. The optimal number of latent states corresponding to the minimum value of BIC is four. The M-T-EM algorithm with parameters $\alpha = 2$ and $\beta = 1.5$ is compared with the EM algorithm according to the same procedure illustrated in Sect. 4.2: for each value of k , 100 different starting values are randomly chosen to initialize both versions of the

Table 14 Maximum log-likelihood and BIC index resulting from fitting a time heterogeneous HM model with EM and O-T-EM algorithms for increasing number of latent states k . For both algorithms, values in bold represent the best results

k	EM		O-T-EM	
	$\hat{\ell}$	BIC	$\hat{\ell}$	BIC
1	-18,100.06	36,339.58	-18,100.06	36,339.58
2	-17,299.80	34,816.53	-17,299.80	34,816.53
3	-16,891.00	34,117.72	-16,887.96	34,111.63
4	-16,386.89	33,269.60	-16,386.89	33,269.60
5	-16,161.01	33,019.26	-16,161.01	33,019.26
6	-16,006.90	32,953.79	-16,002.67	32,945.33
7	-15,859.53	32,943.11	-15,821.86	32,867.78
8	-15,692.55	32,934.54	-15,676.37	32,902.18
9	-15,569.32	33,054.78	-15,531.69	32,979.51
10	-15,459.35	33,242.85	-15,428.07	33,180.30

algorithm. As shown in Table 13, when the chosen HM model is estimated, even in this context, the T-EM guarantees better performance.

More specifically, with the proposed algorithm, the frequency of global maximum is higher: the M-T-EM algorithm reaches the global mode 98 times, while the standard EM algorithm only 73. Moreover, the mean distance from the global optimum decreases to almost zero, and the mean of log-likelihood values increases accordingly; only the median value remains essentially unchanged, with just a very slight enhancement.

5.3 Analyzing countries development

We consider data obtained from the World Bank's World Development Indicators (The World Bank Group 2018) on $n = 175$ countries collected for $T = 5$ years (from 2011 to 2015) on $r = 6$ continuous response variables: life expectancy at birth, total population between the ages 0–14, percentage of population with access to electricity, percentage of population using the internet, share of electricity generated by renewable power plants, and fertility rate. A logit transformation is applied to the variables expressed in a percentage scale, and a Box-Cox transformation (Box and Cox 1964) to all the variables. Results of the estimation of a time heterogeneous HM model on the transformed data with the O-T-EM algorithm for a number of states ranging from 1 to 10 are reported in Table 14. In order to check the assumption on the conditional distribution we check the posterior density of each response variable once the units are allocated according to maximum a posteriori rule; results (available from the authors upon request) seem satisfactory. In this case, the advantages of using the tempering approach are even more evident:

1. it guarantees convergence to the global maximum. Indeed, for most values of k , the maximized log-likelihood value is higher than that of the EM algorithm, showing that the standard EM algorithm cannot correctly detect such a value. Moreover, the mean distance from the global maximum also shows significant improvements,

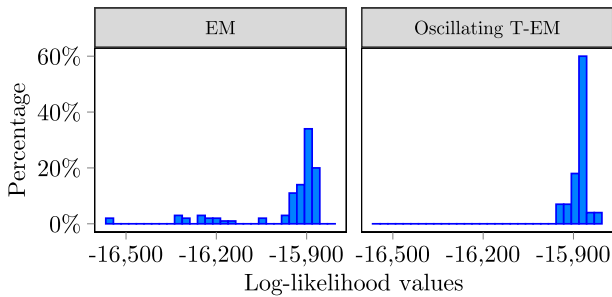


Fig. 6 Maximized log-likelihood values of the HM model for the countries' economic conditions data using standard EM (left) and O-T-EM (right) with $k = 7$ latent states, using 100 random starting values

assuming much smaller values when the O-T-EM algorithm is used, thus showing that it is able to converge repeatedly to the global maximum;

2. it allows us to select a more parsimonious model. Model selection performed with the standard EM leads us to choose eight components, whereas the T-EM algorithm suggests seven components. BIC values are always smaller than those obtained with the standard algorithm;
3. it exhibits an appealing level of flexibility: there is no need to change the optimal set of tempering constants (fixed at $\alpha = 0.6$, $\beta = 110$, $\rho = 5$, and $\tau_0 = 20$) once the HM model is fitted for a number of states ranging from 5 to 10. For values of k from 2 to 4, another unique configuration of tempering constants proves to be the best ($\alpha = 0.5$, $\beta = 120$, $\rho = 5$, and $\tau_0 = 10$).

Focusing on the log-likelihood values shown in Fig. 6 related to the selected model with seven states, we notice that the O-T-EM algorithm always avoids lower values in favor of the higher ones of the maximized log-likelihood. These are reached much more frequently with respect to the EM algorithm.

As already illustrated with the simulation study presented in Sect. 4 and also shown in Table 15, the O-T-EM algorithm is more demanding in terms of computational time with respect to the EM algorithm; however, it has superior performance. Moreover, we notice that on average, a single execution of the T-EM algorithm requires the same time of approximately 10 runs of the standard algorithm. It is important to note that after 1,000 executions performed with 1,000 different random starting values, the EM algorithm is still unable to detect the global maximum (according to the definition provided by the first criterion in Sect. 4) obtained with the O-T-EM algorithm, and equal to $-15,821.86$, since its highest reached value is $-15,834.97$. Neither a higher number of random starting values (up to 10,000 in our study), nor the k -means initialization strategy allows us to improve its performance.

6 Conclusions

The likelihood of discrete latent variable models is typically multimodal, and convergence to a point that it is not the global maximum is a severe limitation of all the algorithms employed for maximum likelihood estimation of the model parameters. To

Table 15 Computational times in seconds of the EM and O-T-EM algorithms. The analysis refers to the estimation of the HM model with continuous response variables for the countries' economic conditions data on the basis of 100 random starting values

Algorithm	Minimum	Median	Mean	Maximum
EM	0.56	2.08	2.33	5.13
O-T-EM	2.03	22.72	28.44	105.69

reduce the chance of local maxima at convergence when the expectation-maximization (EM) algorithm is employed, the model parameters are typically initialized with a multiple-try strategy, employing deterministic and random values. Then, maximum likelihood estimate of the parameters corresponds to the highest log-likelihood at convergence of the algorithm.

In this paper, a new powerful estimation algorithm based on annealing and tempering techniques is proposed in this context. The underlying idea of the tempered EM (T-EM) algorithm is flattening the target function and then gradually warping it back towards the original one. The ability of the algorithm to remain close enough to the dominant maximum is related to the slowness and the graduation of the warping process, which, in turn, is controlled by a sequence of parameters known as the temperature or tempering profile. Two main classes of such profiles usable with many models to be estimated are tested and compared: a monotonically decreasing exponential profile, easy to tune, and an oscillating profile, having more parameters to tune and ensuring best performances with a very high level of flexibility.

An accurate Monte Carlo simulation study is carried out considering two general classes of discrete latent variable models: latent class and hidden Markov models. We compare the performance of the standard EM algorithm with the proposed ones. This comparison is carried out by evaluating the ability to reach the global maximum and the computational time. From the results of the simulation study and those of the applications we show that the proposed algorithms outperform the standard EM, increasing significantly the chance to get to the global maximum in the overwhelming majority of cases. In particular, when an optimally tuned tempering profile is employed, the improvement with respect to the EM algorithm is remarkable: the T-EM algorithm can reach the global mode with a high frequency, generally escaping all local sub-optimal maxima. We detect that the variant with the oscillating profile shows the best performance, slightly outperforming also the monotonic version in most cases.

Estimating the models with the proposed algorithms on categorical and continuous data, having a cross-sectional or longitudinal structure, we also show their good performance in choosing the proper number of latent components. According to the results obtained for the HM model we argue that the proposal may be especially useful for the estimation of the model parameters with complex data structures involving the inclusion of covariates, missing values, and drop-out.

An additional appealing feature of the proposal is the high level of flexibility of the tempering profiles: once a grid-search procedure is employed to set the tempering constants, these constants remain valid also when data with similar characteristics are

used to estimate the model parameters. Moreover, a broad range of values generally performs optimally in many different applied contexts.

Future works may consider the relevant issue of finding a new family of tempering profiles that combine the excellent performance of the oscillating profile with the simple tuning procedure and the fast execution time of the monotonic profile. Other relevant research directions include the exploration of the T-EM algorithm in connection with other maximization algorithms; the most natural choice in this regard is to apply a tempering approach to a direct maximization algorithm, such as Newton-Raphson. The algorithm would also benefit from a more efficient implementation, through the C++ language in order to reduce the computation time. Finally, another possible research line would be to explore and compare the performance of genetic algorithms (Pernkopf and Bouchaffra 2005) with the proposed tempering techniques.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00180-022-01276-7>.

Acknowledgements We greatly acknowledge the Bicocca Data Science Lab (datalab) for supporting this work by providing some computational resources.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix

A Characteristics of the simulated scenarios

Tables 16, 17, and 18 summarize the specific values used to simulate data for the estimation of the LC model, HM model with categorical responses, and HM with continuous responses presented in Sect. 4.1. The following parameters are considered:

- weights of the latent classes (for the LC model) and initial probabilities of the latent states (for the HM models) are defined in such a way that each latent component has the same probability: $\pi_u = 1/k, \forall u = 1, \dots, k$;
- transition probabilities of the HM models are defined to favor persistence in each state; in particular, for $k = 3$ the transition matrix is defined as follows:

$$\begin{bmatrix} 0.800 & 0.150 & 0.050 \\ 0.100 & 0.800 & 0.100 \\ 0.050 & 0.150 & 0.800 \end{bmatrix};$$

Table 16 Description of the scenarios for the LC model: sample size (n), number of response variables (r), categories (c), and latent classes (k)

Scenario	n	r	c	k
A	500	6	3	3
B	1,000	6	3	3
C	500	12	3	3
D	500	6	6	3
E	500	6	3	6

Table 17 Description of the scenarios for the HM model with categorical response variables: sample size (n), number of response variables (r), categories (c), time occasions (T), and latent states (k)

Scenario	n	r	c	T	k
A	500	6	3	5	3
B	1,000	6	3	5	3
C	500	12	3	5	3
D	500	6	6	5	3
E	500	6	3	10	3
F	500	6	3	5	6

Table 18 Description of the scenarios for the HM model with continuous response variables: sample size (n), number of response variables (r), time occasions (T), and latent states (k)

Scenario	n	r	T	k
A	500	6	5	3
B	1,000	6	5	3
C	500	12	5	3
D	500	6	10	3
E	500	6	5	6

- conditional response probabilities are kept fixed considering scenario A (see Tables 16, 17, and 18); for each response variable we define the corresponding matrix as follows:

$$\begin{bmatrix} 0.800 & 0.100 & 0.050 \\ 0.150 & 0.800 & 0.150 \\ 0.050 & 0.100 & 0.800 \end{bmatrix};$$

- for the HM model with continuous response variables, the same conditional distribution holds for all response variables; for example, with $k = 3$ latent states, the mean vector $\mu = [-2, 0, 2]'$ is fixed for each response variable;
- the variance-covariance matrix Σ is computed as the sample covariance matrix of the data.

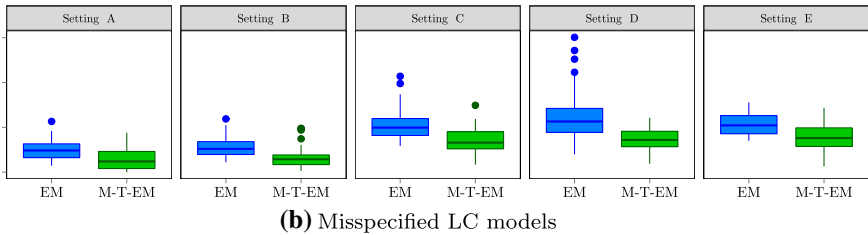
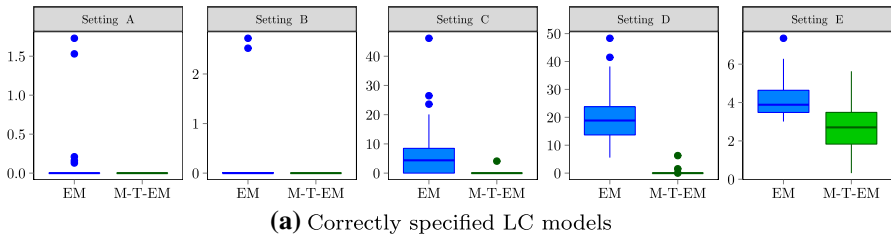


Fig. 7 Mean distance from the global maximum using EM and M-T-EM algorithms under the simulated scenarios presented in Table 16 of the Appendix A for the LC model

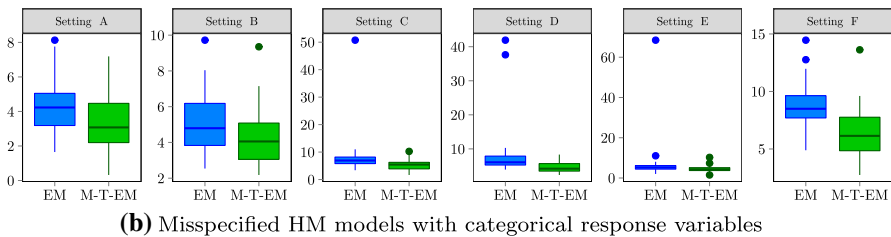
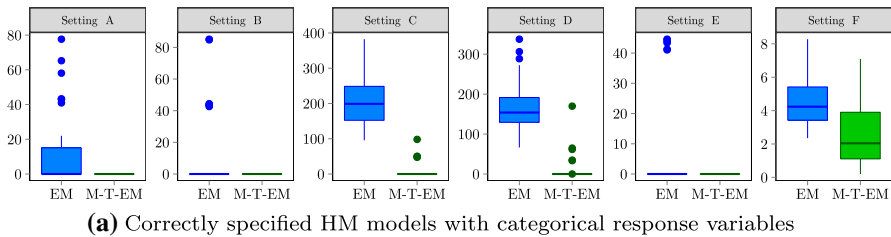
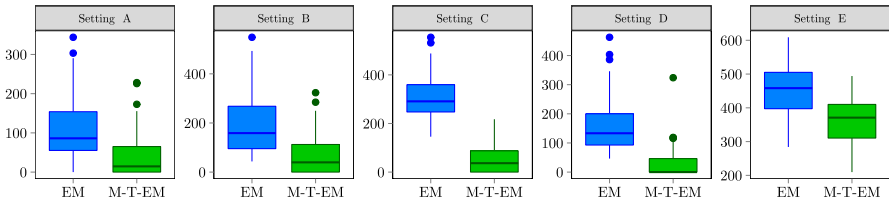


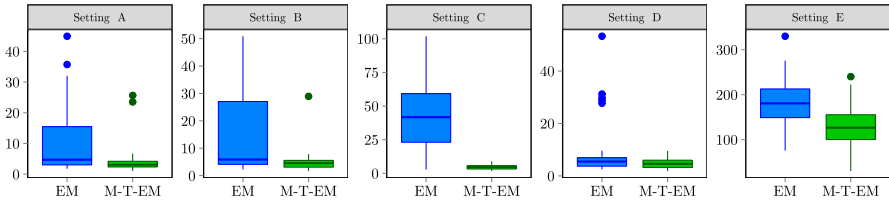
Fig. 8 Mean distance from the global maximum using EM and M-T-EM algorithms under the simulated scenarios presented in Table 17 of the Appendix A for the HM model with categorical response variables

B Additional simulation results

In this section we report additional details on the results of the simulation study in Sect. 4.2. In particular, for each considered scenario (see Tables 16, 17, and 18), Figs. 7, 8, and 9 show the distribution of the mean distance from the global maximum through boxplots.



(a) Correctly specified HM models with continuous response variables



(b) Misspecified HM models with continuous response variables

Fig. 9 Mean distance from the global maximum using EM and M-T-EM algorithms under the simulated scenarios presented in Table 18 of the Appendix A for the HM model with continuous response variables

C Numerical results for the analysis of T-EM algorithm with fixed tempering profiles

In this section we present results obtained from the simulation studies comparing the EM algorithm with the T-EM algorithm with fixed tempering profiles; the analysis carried out on the basis of the results is reported in Sect. 4.6. See Tables 19, 20, and 21.

Table 19 Performance of the M-T-EM algorithm under simulated scenarios presented in Table 16 in the Appendix A for the LC model using fixed configurations of tempering constants α and β . The last column shows the percentage of samples for which the M-T-EM algorithm outperforms the EM algorithm











Scenario	Tempering profile		Percentage
	α	β	
<i>Correctly specified LC model</i>			
A	6	0.7	58% 
B	6	0.6	62% 
C	1	0.7	78% 
D	1	1.8	72% 
E	1	1.5	64% 
<i>Misspecified LC model</i>			
A	2	0.6	50% 
B	2	0.6	52% 
C	2	0.6	44% 
D	2	0.6	58% 
E	2	0.6	62% 

Table 20 Performance of the M-T-EM algorithm under simulated scenarios presented in Table 16 in the Appendix A for the HM model with categorical response variables using fixed configurations of tempering constants α and β . The last column shows the percentage of samples for which the M-T-EM algorithm outperforms the EM algorithm























Scenario	Tempering profile		Percentage
	α	β	
<i>Correctly specified categorical HM model</i>			
A	1	0.7	96% 
B	1	0.6	100% 
C	1	1.9	74% 
D	3	1.8	70% 
E	1	0.6	100% 
F	14	1.1	64% 
<i>Misspecified categorical HM model</i>			
A	5	2.0	60% 
B	5	1.9	54% 
C	6	0.0	66% 
D	3	1.7	64% 
E	1	1.9	52% 
F	15	0.1	42% 

Table 21 Performance of the M-T-EM algorithm under simulated scenarios presented in Table 16 in the Appendix A for the HM model with continuous response variables using fixed configurations of tempering constants α and β . The last column shows the percentage of samples for which the M-T-EM algorithm outperforms the EM algorithm

Scenario	Tempering profile		Percentage
	α	β	
Correctly specified continuous HM model			
A	1	1.3	92% 
B	1	1.1	100% 
C	2	0.3	96% 
D	2	0.2	98% 
E	1	1.1	100% 
Misspecified continuous HM model			
A	2	0.4	84% 
B	1	0.9	92% 
C	1	1.2	80% 
D	2	0.2	80% 
E	3	0.0	86% 

References

- Barbu A, Zhu S (2013) Monte Carlo methods. Springer, Singapore
- Bartolucci F, Bacci S, Gnaldi M (2014) `MultiLCIRT`: an R package for multidimensional latent class item response models. *Comput Stat Data Anal* 71:971–985
- Bartolucci F, Farcomeni A, Pennoni F (2013) Latent Markov models for longitudinal data. Chapman and Hall/CRC, Boca Raton
- 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* 23:433–486
- Bartolucci F, Pandolfi S, Pennoni F (2017) `LMest`: an R package for latent Markov models for longitudinal categorical data. *J Stat Softw* 81:1–38
- Bartolucci F, Pandolfi S, Pennoni F (2022) Discrete latent variable models. *Annu Rev Stat Appl* 6:1–31
- Bartolucci F, Pennoni F, Francis B (2007) A latent Markov model for detecting patterns of criminal activity. *J R Stat Soc Ser A Stat Soc* 170:114–132
- Baum L, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat* 41:164–171
- Box GE, Cox DR (1964) An analysis of transformations. *J R Stat Soc Ser B Stat Methodol* 26:211–243
- Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc Ser B Stat Methodol* 39:1–38
- Earl DJ, Deem MW (2005) Parallel tempering: theory, applications, and new perspectives. *Phys Chem Chem Phys* 7:3910–3916
- Everitt BS, Landau S, Leese M, Stahl D (2011) Cluster analysis, 5th edn. Wiley, New York
- Falcioni M, Deem M (1999) A biased Monte Carlo scheme for zeolite structure solution. *J Chem Phys* 110:1754–1766
- Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood. In: Computing science and statistics, proceedings of the 23rd symposium on the interface, computing science and statistics. Interface Foundation of North America, pp 156–163
- Geyer CJ, Thompson EA (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. *J Am Stat Assoc* 90:909–920
- Goodman L (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215–231
- Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their application. *Biometrika* 57:97–109

- Hofmann CJ (1999) Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence, UAI'99. Morgan Kaufmann Publisher Inc., San Francisco, CA, USA, pp 289–296
- Huang Z (1998) Extensions to the k -means algorithm for clustering large data sets with categorical values. *Data Min Knowl Discov* 2:283–304
- Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. *Science* 220:671–680
- Lartigue T, Durrleman S, Allasonnière S (2022) Deterministic approximate EM algorithm; application to the Riemann approximation EM and the tempered EM. *Algorithms* 15:78
- Lazarsfeld P, Henry N (1968) Latent structure analysis. Houghton Mifflin, Boston
- Leroux B, Puterman M (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* 48:545–558
- Lindsay B, Clogg C, Grego J (1991) Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J Am Stat Assoc* 86:96–107
- MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, UAI'99. University of California Press, Berkeley, CA, USA, pp 281–297
- Maruotti A, Punzo A (2021) Initialization of hidden Markov and semi-hidden Markov: a critical evaluation of several strategies. *Int Stat Rev* 89:447–480
- McLachlan G, Basford K (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York
- McLachlan G, Krishnan T (2008) The EM algorithm and extensions, 2nd edn. Wiley, Hoboken
- Metropolis N, Rosenbluth A, Rosenbluth M, Teller A-H, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092
- Pandolfi S, Bartolucci F, Pennoni F (2021) Maximum likelihood estimation of hidden Markov models for continuous longitudinal data with missing responses and dropout. [arXiv:2106.15948](https://arxiv.org/abs/2106.15948), 1–36
- Pennoni F (2014) Issues on the estimation of latent variable and latent class models. Scholar's Press, Saarbrücken
- Pernkopf F, Bouchaffra D (2005) Genetic-based em algorithm for learning gaussian mixture models. *IEEE Trans Pattern Anal Mach Intell* 27:1344–1348
- R Core Team (2022) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Research Development and Statistics Directorate (1998) The offenders index: codebook. <https://homeoffice.gov.uk/rds/pdf/soicodes.pdf>
- Robert C, Elvira V, Tawn N, Wu C (2018) Accelerating MCMC algorithms. *Wiley Interdiscip Rev Comput Stat* 10:1–14
- Sambridge M (2014) A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophys J Int* 196:357–374
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- The World Bank Group (2018). Data catalog: World development indicators. <https://datacatalog.worldbank.org/dataset/world-development-indicators>
- Ueda N, Nakano R (1998) Deterministic annealing EM algorithm. *Neural Netw* 11:271–282
- Yuille A, Stolorz P, Utans J (1994) Statistical physics, mixture of distributions, and the EM algorithm. *Neural Comput* 6:334–340
- Zhou H, Lange K (2010) On the bumpy road to the dominant mode. *Scand J Stat* 37:612–631
- Zigmond A, Snaith R (1983) The hospital anxiety and depression scale. *Acta Psychiatr Scand* 67:361–70
- Zucchini W, Guttorp P (1991) A hidden Markov model for space-time precipitation. *Water Resour Res* 27:1917–1923
- Zucchini W, MacDonald I, Langrock R (2016) Hidden Markov models for time series: an introduction using R, 2nd edn. Chapman & Hall/CRC, Boca Raton

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.