Book of Short Papers SIS 2021

Editors: Cira Perna, Nicola Salvati and Francesco Schirripa Spagnolo

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891927361

Contents

A Tempered Expectation-Maximization Algorithm for Latent Class Model Estimation

Un Algoritmo Tempered Expectation-Maximization per la Stima del Modello a Classi Latenti

Luca Brusa, Francesco Bartolucci and Fulvia Pennoni

Abstract We consider maximum likelihood estimation of the Latent Class model, which is formulated through individual discrete latent variables. We explore tempering techniques to overcome the problem of multimodality of the log-likelihood function. A Tempered Expectation-Maximization algorithm is proposed, which can adequately explore the parameter space and reach the global maximum more frequently than the standard EM algorithm. We assess the performance of the proposed approach by a Monte Carlo simulation study and an application based on data about anxiety and depression in oncological patients.

Abstract *Consideriamo la stima di massima verosimiglianza del modello a classi latenti che è formulato attraverso variabili latenti discrete a livello individuale. Esploriamo le tecniche di tempering per fronteggiare il problema della multimodalità della funzione di log-verosimiglianza. Proponiamo un algoritmo denominato Tempered Expectation-Maximization che permette di esplorare adeguatamente lo spazio dei parametri e di raggiungere il massimo globale più frequentemente rispetto all'usuale algoritmo EM. Per valutare l'efficacia della proposta utilizziamo uno studio di simulazione Monte Carlo e un'applicazione basata su dati reali riguardanti misure di ansia e depressione in pazienti oncologici.*

Key words: annealing, finite mixture models, latent variables, local maxima.

Luca Brusa,
Department of Economics, Management and Statistics, University of Milano-Bicocca,
e-mail: luca.brusa@unimib.it

Francesco Bartolucci,
Department of Economics, University of Perugia,
e-mail: francesco.bartolucci@unipg.it

Fulvia Pennoni,
Department of Statistics and Quantitative Methods, University of Milano-Bicocca,
e-mail: fulvia.pennoni@unimib.it

1 Introduction

The Latent Class (LC) model [1] is very popular for the analysis of categorical, and in particular binary, response variables. It is formulated by assuming the existence of individual-specific latent variables having a discrete distribution. This model may be seen as semi-parametric since, differently from other models based on continuous latent variables, no parametric assumptions are formulated on the distribution of such variables. The LC model may be seen as a finite mixture model, and it is employed to cluster subjects on the basis of a set of categorical, typically binary, responses.

Despite maximum likelihood estimation of the LC model may be simply performed using the Expectation-Maximization (EM) algorithm [2, 3], a well-known drawback of this estimation method is related to the multimodality of the likelihood function that is due to the inclusion of discrete latent variables. The consequence is that the *global* maximum of the likelihood is not ensured to be reached, and a proper initialization of the estimation algorithm is crucial. A multi-start strategy is typically adopted based on deterministic and random rules to explore the parameter space adequately. However, this approach may be computationally intensive, and it does not guarantee convergence to the global maximum.

In order to face the multimodality of the likelihood function, we propose a Tempered EM (T-EM) algorithm able to explore the parameter space adequately. In an optimization context, tempering [4] consists of re-scaling the objective function depending on a variable, known as temperature, which controls the prominence of global and local maxima. High temperatures allow us to explore wide regions of the parameter space, avoiding the maximization algorithm being trapped in non-global maxima; low temperatures, instead, guarantee a sharp optimization in a local region of the parameter space. By properly tuning the sequence of temperature values, the procedure is gradually attracted toward the global maximum, escaping in this way local sub-optimal solutions. As a future development, this procedure will also be applied to estimate the parameters of the hidden Markov (HM) models for the analysis of longitudinal data [5].

The rest of the paper is organized as follows. Section 2 outlines the LC model formulation and maximum likelihood estimation through the EM algorithm. Section 3 provides details on the proposed T-EM algorithm. Section 4 summarizes the main findings of the simulation study and the results of an application concerning patients' responses to ordinal items measuring anxiety and depression.

2 Latent Class Model and Expectation-Maximization Algorithm

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ir})'$ denote the vector of r categorical response variables for individual $i = 1, \ldots, n$; each variable has the same number c of categories, labeled from 0 to $c-1$. The LC model relies on individual-specific discrete latent variables U_i with k support points that identify the latent classes in the population. The model parameters are the conditional probabilities of each response variable given the latent variable, denoted by $\phi_{jy|u} = p(Y_{ij} = y|U_i = u)$, and the weight of each latent class, denoted by $\pi_u = p(U_i = u)$. The resulting *manifest distribution* is then

$$p(\boldsymbol{y}_i) = \sum_{u=1}^{k} \pi_u \prod_{j=1}^{r} \phi_{jy_{ij}|u},$$

where $\boldsymbol{y}_i$ denotes a realization of $\boldsymbol{Y}_i$.

In order to estimate the model parameters, collected in the vector $\boldsymbol{\theta}$, on the basis of a sample of n independent observations $\boldsymbol{y}_i$, we rely on the log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i).$$

This function is maximized through the EM algorithm on the basis of the *complete data log-likelihood*, which may be written as

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

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 responded 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. The EM algorithm alternates the following two steps until a suitable convergence criterion is satisfied:

- **E-Step**: compute the conditional expected value of $\ell^*(\boldsymbol{\theta})$, given the observed data and the value of the parameters at the previous step;
- **M-Step**: maximize the expected value of the log-likelihood function $\ell^*(\boldsymbol{\theta})$ and so update the model parameters.

In particular, the E-step is based on the posterior probabilities

$$q(u|\boldsymbol{y}_i) = p(U_i = u|\boldsymbol{Y}_i = \boldsymbol{y}_i) = \frac{\pi_u \prod_{j=1}^{r} \phi_{jy_{ij}|u}}{p(\boldsymbol{y}_i)},$$

on the basis of which the expected values of the frequencies a_{juy} and b_u are simply obtained.

The EM algorithm is straightforward to implement, it is able to converge in a stable way to a local maximum of the log-likelihood function, and it is used for parameter estimation in many available packages [3]. However, this log-likelihood function may be multimodal, especially when the model has many latent classes. For this reason, several starting values of the parameters in $\boldsymbol{\theta}$ are typically used, and the solution corresponding to the highest log-likelihood is then selected as the maximum likelihood estimate, denoted by $\hat{\boldsymbol{\theta}}$. In the next section, we show an alternative solution based on the proposed T-EM algorithm.

3 Tempered Expectation-Maximization Algorithm

We introduce a T-EM algorithm [6, 7] by defining the following modified posterior probabilities:

$$\tilde{q}^{(T_h)}(u|\boldsymbol{y}_i) \doteq \frac{q(u|\boldsymbol{y}_i)^{1/T_h}}{\sum_{u=1}^{k} q(u|\boldsymbol{y}_i)^{1/T_h}},$$

where $(T_h)_{h\geq 1}$ is a suitable sequence of temperature values, under the constraint that $T_h \to 1$ as $h \to \infty$, where h is the algorithm iteration number.

The E-step and M-step of the T-EM algorithm are implemented as follows by modifying those of the original EM algorithm:

- **E-Step**: compute

$$\tilde{b}_u^{(T_h)} = \sum_{i=1}^{n} \tilde{q}^{(T_h)}(u|\boldsymbol{y}_i) \qquad \text{and} \qquad \tilde{a}_{juy}^{(T_h)} = \sum_{i=1}^{n} I(y_{ij} = y)\, \tilde{q}^{(T_h)}(u|\boldsymbol{y}_i);$$

- **M-Step**: update the parameters as

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

Given the above setting, it is clear that the tempering profile (i.e., the sequence $(T_h)_{h\geq 1}$) may have a deep impact on the performance of the proposed algorithm. In fact, increasing the temperature value has the effect of flattening the profile of the log-likelihood, thereby reducing the chance that the algorithm will get trapped into local maxima. In particular, $T_h \to +\infty$ yields $\tilde{q}^{(T_h)}(u|\boldsymbol{y}_i)$ to a uniform distribution, while $T_h = 1$ makes $\tilde{q}^{(T_h)}(u|\boldsymbol{y}_i)$ equal to the standard posterior probability $q(u|\boldsymbol{y}_i)$. Therefore, the only necessary condition for proper convergence is that the temperature value T_h tends towards 1 as the iteration counter increases.

We consider the following two tempering profiles: (*i*) a decreasing exponential profile:

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

with constants $\alpha \geq 1$ and $\beta \geq 0$, which has the advantage to be easy to tune; (*ii*) a non-monotonic profile [6] with oscillations of gradually smaller amplitude:

$$T_h = \tanh\left(\frac{h}{2r}\right) + \left(T_0 - \beta \cdot \frac{2\sqrt{2}}{3\pi}\right) \cdot \alpha^{h/r} + \beta \cdot \operatorname{sinc}\left(\frac{3\pi}{4} + \frac{h}{r}\right), \tag{2}$$

with constants $r,\ T_0,\ \beta > 0$, and $0 < \alpha < 1$. The latter choice has more parameters to tune, but it guarantees a very high level of flexibility. The proposed procedure requires selecting the set of tempering parameters by a grid-search.

4 Simulation Results and Applicative Example

Within the simulation study, we randomly drew several samples of size $n = 500$ from an LC model with $r = 6$ responses, having $c = 3$ categories, assuming $k = 3$ latent classes, and for each of these samples we estimated a misspecified LC model with $k = 4$ latent classes. In particular, we fitted the LC model 100 times for each sample, always using different sets of random starting values. We used the standard EM algorithm and the proposed T-EM algorithms denoted by M. T-EM, when the monotonic tempering profile (1) is used, and by O. T-EM, when the oscillating tempering profile (2) is employed. The convergence of the algorithms is checked on the basis of the relative log-likelihood difference; regarding the algorithm initialization, we adopted a random starting rule based on normalized random numbers drawn from a uniform distribution from 0 to 1.

We carried out a grid-search for the tempering parameters for each sample, and we evaluated the setting that ensures the best performance. We noticed that the method is not excessively sensitive to the tempering parameters: once the grid-search sets such parameters, they remain valid over datasets sharing the same features (e.g., the same number of response variables and categories). Therefore, this preliminary procedure may be less time consuming than the current practice of estimating the model many times with random initial parameters.

In Table 1 we show some results obtained as described above about the EM and T-EM algorithms: for each of six considered samples, we report the mean and the median of the 100 log-likelihood values at convergence. From this table, it is clear the advantage of the use of the tempering modification. In particular, the oscillating version of the T-EM algorithm exhibits the best performance, slightly outperforming also the monotonic version in most cases. We also considered the following criteria: (*i*) dispersion of the resulting maxima measured by the standard deviation; (*ii*) proportion of times the obtained maximum is close enough to the global one (based on the 100 repetitions); (*iii*) dispersion of the estimated probability vectors ($\boldsymbol{\pi}$, sorted into descending order). From the results reported in Table 2 we notice a clear superiority of T-EM algorithm over the standard EM algorithm: in each scenario the best results are obtained with the modified algorithm, and only for the fourth sample the improvement is mild.

Table 1 Mean and median of log-likelihood values at the maximum, with EM and T-EM algorithm using monotonic (M. T-EM) and oscillating (O. T-EM) tempering profiles on simulated data; each row refers to a specific sample, and values in bold highlight the best results.

Mean			Median		
EM	M. T-EM	O. T-EM	EM	M. T-EM	O. T-EM
-2,847.2879	-2,846.5392	**-2,845.4207**	-2,846.7726	-2,844.9000	**-2,844.8369**
-2,864.7102	-2,864.8438	**-2,864.6754**	-2,864.8575	-2,864.7875	**-2,864.7336**
-2,848.3929	-2,846.4819	**-2,846.4817**	-2,849.0982	-2,846.4819	**-2,846.4817**
-2,798.8988	-2,798.7510	**-2,798.3810**	-2,799.5792	-2,797.9326	**-2,797.5355**
-2,846.4159	**-2,843.4433**	-2,843.4672	-2,847.5666	**-2,843.4433**	-2,843.4449
-2,832.5526	**-2,831.5140**	-2,831.5808	-2,831.9158	**-2,831.2970**	**-2,831.2970**

Table 2 Dispersion (SD) and proportion (Freq) of global maxima and dispersion of the estimated probabilities ($\boldsymbol{\pi}$); each row refers to a specific sample, and values in bold highlight the best results.

SD (Max.)			Freq. (Glob. Max.)			Dispersion of $\boldsymbol{\pi}$		
EM	M. T-EM	O. T-EM	EM	M. T-EM	O. T-EM	EM	M. T-EM	O. T-EM
2.2106	2.0383	**1.3565**	0.63	0.67	**0.91**	0.0057	0.0042	**0.0009**
1.1132	**0.7430**	1.0831	0.89	**0.93**	0.85	0.0029	0.0024	**0.0023**
1.9395	**0.0000**	0.0000	0.64	**1.00**	**1.00**	0.0067	**0.0000**	0.0000
1.7738	1.6139	**1.5742**	0.49	0.51	**0.62**	0.0038	0.0033	**0.0029**
2.8364	**0.0000**	0.0298	0.47	**1.00**	**1.00**	0.0024	**0.0000**	0.0002
1.5343	0.3525	**0.3404**	0.88	**1.00**	**1.00**	0.0043	0.0005	**0.0004**

Comparing the two types of tempering profile, we note that the oscillating profile often outperforms the monotonic one; only when the results exhibit an almost absolute perfection (dispersion approximately equal to 0 and proportion of global maxima close to 1, as in samples 3 and 5), the monotonic profile reaches a slightly better performance. This observation suggests that if the model is not too complex, this choice is generally preferable, while in other cases, the oscillating profile guarantees better results. Similar T-EM algorithms are implemented for estimating the HM model and preliminary results of the simulation study show the same improvements with respect to the standard EM algorithm.

We also considered data deriving from the administration of 14 ordinal items with three categories measuring anxiety and depression in 201 oncological patients [3]. A misspecified LC model is estimated with $k = 4$ latent classes with the following tempering parameters: $\alpha = 42$ and $\beta = 1.5$ for the monotonic profile; $r = 90$, $T_0 = 10$, $\beta = 20$, and $\alpha = 0.8$ for the oscillating profile. We notice that the T-EM algorithms outperform the classic version of the EM algorithm because they always converge to the same value that is presumably the global maximum, while the classical EM algorithm spreads out over a wide range of estimates.

References

1. Goodman, L.A.: Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biom.*, **61**, 215-231 (1974).
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from in-complete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B*, **39**, 1–38 (1977).
3. Bartolucci, F., Bacci, S., Gnaldi, M.: `MultiLCIRT`: An `R` package for multidimensional latent class item response models. *Comput. Stat. Data Anal.*, **71**, 971–985 (2014).
4. Sambridge, M.: A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophys. J. Int.*, **196**, 357–374 (2013).
5. Bartolucci, F., Farcomeni, A., Pennoni, F.: *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, Boca Raton (2013).
6. Lartigue, T., Durrleman, S., Allassonnière, S.: Deterministic approximate EM algorithm; Application to the Riemann approximation EM and the tempered EM. *arXiv:2003.10126* (2021).
7. Ueda, N., Nakano, R.: Deterministic annealing EM algorithm. *Neural Netw.*, **11**, 271–282 (1998).