Tempered Expectation-Maximization algorithm for discrete latent variable models

Luca Brusa *(luca.brusa@unimib.it)*

University of Milano-Bicocca - Department of Economics, Management and Statistics

Francesco Bartolucci *(francesco.bartolucci@unipg.it)*

University of Perugia - Department of Economics

Fulvia Pennoni *(fulvia.pennoni@unimib.it)*

University of Milano-Bicocca - Department of Statistics, and Quantitative Methods

February 22, 2022

Overview

Latent variable models

- A **latent variable model** (Bartolucci et al., 2022) is a statistical model in which the distribution of the response variables is affected by one or more variables that are not directly observable

- A possible classification of these models distinguishes between discrete and continuous latent variables; here, we consider two special classes of discrete latent variable models, namely latent class and hidden Markov

Expectation-Maximization algorithm

- Maximum likelihood estimation of model parameters is based on the complete data log-likelihood function $\ell^*(\boldsymbol{\theta})$ and it is performed through the **Expectation-Maximization** (**EM**) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008)
- Alternate the following steps until a suitable convergence condition:
  - **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 $\ell^*(\boldsymbol{\theta})$ and so update the model parameters
- The E-step is based on specific conditional posterior probabilities, with respect to which the expected values are computed, in the following generically referred to as $q(\cdot)$

The problem of local maxima

- 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
- However, a well-known drawback is related to the **multimodality** of the log-likelihood function, especially when the model has many latent classes; therefore the global maximum is not ensured to be reached
- Currently, a **multi-start strategy** is typically adopted, based on deterministic and random rules; however this approach may be computationally intensive and it does not guarantee convergence to the global maximum

Tempering approach

- In an optimization context, **tempering** and **annealing** (Sambridge, 2013), constitute a broad family of methods consisting in re-scaling the objective function on the basis of a variable, known as **temperature**, that controls the prominence of all maxima
- By properly **tuning the sequence of temperature** values, the procedure is gradually attracted towards the global maximum, escaping local sub-optimal solutions:
  - **high temperatures** allow exploring wide regions of the parameter space, avoiding being trapped in non-global maxima
  - **low temperatures** guarantee a sharp optimization in a local region of the solution space
- A similar approach was applied to Gaussian mixture models (Ueda and Nakano, 1998; Zhou and Lange, 2010; Lartigue et al., 2021)

Tempered EM Algorithm: Derivation

- We implement the **tempered EM** (**T-EM**) algorithm, by adjusting the computation of the conditional expected frequencies in the E-step
- We define the following family of **tempered probabilities**:

$$\tilde{q}^{(\tau)}(\cdot) \propto q(\cdot)^{1/\tau},$$

  where $\tau \in [1, +\infty)$ is the temperature value such that:
  - the choice $\tau \to +\infty$ yields $\tilde{q}^{(\tau)}(\cdot)$ to a uniform distribution
  - the choice $\tau = 1$ recovers the original distribution $q(\cdot)$
- At each E-step of the T-EM algorithm, the conditional expected frequencies are computed accordingly

Tempering profile

- We define a sequence of temperatures $(\tau_h)_{h \geq 1}$, such that:
  - τ_1 is sufficiently small so that $\tilde{q}^{(\tau_1)}(\cdot)$ is relatively flat
  - τ_h tends towards 1 as the algorithm iteration counter increases
- The resulting sequence, known as **tempering profile**, guarantees a proper convergence of the T-EM algorithm
- To ensure flexibility to the tempering profile, it depends on a set of constants; a suitable grid-search procedure is employed to select the optimal configuration of tempering constants

Tempering profile

In particular, we consider two classes of tempering profiles:

- a **monotonically decreasing exponential** (M-T-EM) profile:

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

- a **non-monotonic** profile with **gradually smaller oscillations** (O-T-EM)

$$\tau_h = \tanh\left(\frac{h}{2\rho}\right) + \left(T_0 - \beta\,\frac{2\sqrt{2}}{3\pi}\right)\alpha^{h/\rho} + \beta\,\mathrm{sinc}\left(\frac{3\pi}{4} + \frac{h}{\rho}\right)$$

Simulation study

- To evaluate the performance of the T-EM algorithm, we conduct an extensive **Monte Carlo simulation study**:
  1. after fixing a set of models parameters, we draw many samples from the corresponding model
  2. for each sample, we estimate a misspecified model 100 times; in particular, 100 starting values are randomly selected and employed to fit the model with EM, M-T-EM and O-T-EM algorithms
  3. on the basis of the maximized log-likelihood values, we compare the performance of the EM and T-EM algorithms

- The criteria considered to compare the behavior of the algorithms are the following:
  - **mean** and **median** of the maximized log-likelihood values
  - **ability to reach the global maximum**
  - **mean distance** from the global maximum

Objectives of the study

1. We analyze the **performance** of the T-EM algorithm when the termpering profile is **optimally tuned** through a grid-search procedure
2. We test the proposal **without** performing a preliminary **tuning** procedure for the tempering constants, but **fixing** them in advance
3. We compare the proposed T-EM algorithm with the original EM algorithm with regard to the **computational time**

1. Analysis of T-EM with optimally tuned profiles

- Employing the **T-EM** algorithm, the distribution of maximum log-likelihood values appears to be much more **concentrated** towards the **global maximum**: both mean and median show significantly higher values with the M-T-EM or O-T-EM
- With the standard **EM** algorithm the proportion of times the global maximum is reached rarely exceeds **70%**. With the **T-EM** algorithms, such a proportion noticeably increases: it results very often equal to **100%**, thus meaning that the algorithm always leads to the global maximum
- When the **T-EM** algorithm is employed, the **mean distance** from the global maximum **decreases** for all samples, often reaching very low values. Only in a tiny minority of cases this improvement is just mild

1. Analysis of T-EM with optimally tuned profiles

- As an example, we propose a portion of the results obtained for the **LC model**
- We compare the mean and the median of maximized log-likelihood values, proportion of global maximum and mean distance from the global mode, using EM, M-T-EM and O-T-EM alogorithms; each row refers to a specific sample, and values in bold highlight the best results

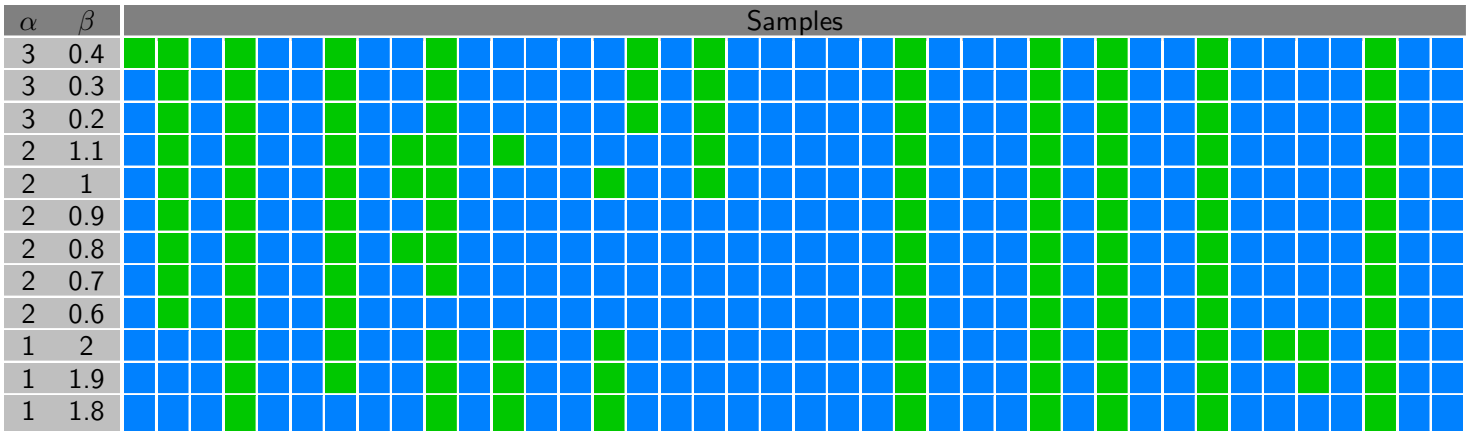
Mean			Median			Freq. [Dist.] (Glob. Max)		
EM	M-T-EM	O-T-EM	EM	M-T-EM	O-T-EM	EM	M-T-EM	O-T-EM
-2821.79	-2820.59	**-2820.20**	-2820.80	-2820.20	**-2820.20**	71% [1.59]	91% [0.39]	**100%** [**0.00**]
-2859.49	-2859.17	**-2858.32**	-2859.97	-2858.32	**-2858.32**	81% [1.75]	94% [1.43]	**100%** [**0.58**]
-2816.34	-2813.95	**-2813.84**	-2813.99	-2813.99	**-2813.99**	74% [3.26]	99% [0.87]	**100%** [**0.76**]
-2771.62	-2769.18	**-2768.82**	-2768.82	-2768.82	**-2768.82**	62% [2.80]	94% [0.36]	**100%** [**0.00**]
-2834.76	-2833.62	**-2833.19**	-2833.19	-2833.19	**-2833.19**	64% [1.57]	89% [0.43]	**100%** [**0.00**]
-2841.04	-2840.85	**-2840.46**	-2840.95	-2840.95	**-2840.95**	93% [1.60]	**100%** [1.41]	**100%** [**1.02**]
-2807.84	**-2806.24**	-2806.68	-2807.49	**-2805.89**	-2806.68	69% [2.42]	**100%** [**0.82**]	**100%** [1.26]
-2802.44	-2800.01	**-2799.58**	-2801.32	-2799.58	**-2799.58**	66% [2.86]	95% [0.43]	**100%** [**0.00**]
-2880.24	-2879.76	**-2879.06**	-2879.10	-2879.07	**-2879.06**	67% [1.65]	76% [1.16]	**100%** [**0.47**]
-2846.62	-2845.37	**-2845.07**	-2845.14	-2845.07	**-2845.07**	66% [2.56]	95% [1.30]	**100%** [**1.01**]

2. Analysis of T-EM with fixed profiles

- All the chosen configurations of tempering parameters provide excellent results: given a fixed configuration, the **T-EM** algorithm **outmatches** the standard version in around **70% of samples**
- Once a **configuration** of tempering constants is set by grid-search over a specific sample, it generically **remains valid** for more than 70% of other samples sharing the same features (mainly, the same number of response variables and categories)
- For most cases, the optimal **configurations** of tempering constants have to be chosen from a list simply depending on the **sample characteristics**, thus making the tuning procedure significantly faster, and the proposed tempering approach suitable and sufficiently general for a broad class of models
- Only a **few samples** still require a **proper and complete tuning** of tempering profile

2. Analysis of T-EM with fixed profiles

- The advantage is relevant for both the monotonic and the oscillating tempering profiles and for both models
- In the example below, a list of 12 different configurations of tempering constants (on the rows) is considered for applying the M-T-EM algorithm to 40 different samples (on the columns) drawn from an **HM** model with categorical responses
- When the M-T-EM version outperforms the standard EM algorithm in all criteria, a blue square is shown; when at least one criterion shows a better result for the classic EM algorithm, a green square is shown



3. Analysis of T-EM in terms of computational time

Computational time (in seconds) of the EM, M-T-EM, and O-T-EM algorithms considering 50 samples drawn from the **LC** model and 100 starting values for each sample

Algorithm	Minimum	Median	Mean	Maximum
EM	0.0571	0.3050	0.3591	1.9521
M-T-EM	0.1025	0.4727	0.5647	2.1454
O-T-EM	0.1023	28.5897	28.9212	63.4069

- **EM** algorithm and **M-T-EM** are approximately **equally fast**
- **O-T-EM** requires a much **larger computational time** (up to 30 times slower)

3. Analysis of T-EM in terms of computational time

Computational time (in seconds) of the EM and M-T-EM algorithm considering 40 samples drawn from the **HM** model with categorical responses and 100 starting values for each sample

Algorithm	Minimum	Median	Mean	Maximum
EM	0.098	2.319	2.705	13.332
M-T-EM	3.261	26.502	31.635	109.250

- Differently from what observed for the LC model, the computational time of the **M-T-EM** algorithm is **higher** than that of the standard **EM** algorithm
- The overall behavior of the **T-EM** algorithm is still the **best**, since a single execution requires the same time of about 10 runs of the EM algorithm, which are insufficient to detect the global maximum

References I

Bartolucci, F., Bacci, S., and Gnaldi, M. (2014). `MultiLCIRT`: an `R` package for multidimensional latent class item response models. *Computational Statistics and Data Analysis*, **71**, 971–985.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov models for longitudinal data*. Chapman & Hall/CRC, Boca Raton.

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). `LMest`: an `R` package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, **81**, 1–38.

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2022). Discrete latent variable models. *Annual Review of Statistics and its Application*, **6**, 1–31.

References II

Brusa, L., Bartolucci, F., and Pennoni, F. (To be submitted). Tempered Expectation-Maximization algorithm for discrete latent variable models.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

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

Lartigue, T., Durrleman, S., and Allassonnière, S. (2021). Deterministic approximate EM algorithm; application to the Riemann approximation EM and the tempered EM.

McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions: 2nd edition*. John Wiley and Sons, Hoboken, NJ.

References III

Sambridge, M. (2013). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophys. J. Int.*, **196**, 357–374.

Ueda, N. and Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Netw.*, **11**, 271–282.

Zhou, H. and Lange, K. (2010). On the bumpy road to the dominant mode. *Scand. J. Stat.*, **37**, 612–631.

Zucchini, W., MacDonald, I., and Langrock, R. (2016). *Hidden Markov models for time series: an introduction using R, 2nd edition*. Chapman & Hall/CRC, Boca Raton.