Statistics and Probability Letters 146 (2019) 97–103

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A nested expectation–maximization algorithm for latent class models with covariates

Daniele Durante [a,*], Antonio Canale [b], Tommaso Rigon [a]

[a] Department of Decision Sciences, Bocconi University, Via Röntgen 1, 20136 Milan, Italy
[b] Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy

ARTICLE INFO

Article history:
Received 16 January 2018
Received in revised form 22 August 2018
Accepted 25 October 2018
Available online 9 November 2018

Keywords:
EM algorithm
Latent class model
Multivariate categorical data
Pólya-gamma

ABSTRACT

We propose a nested EM routine which guarantees monotone log-likelihood sequences and improved convergence rates in maximum likelihood estimation of latent class models with covariates.

1. Introduction

Multivariate categorical data are routinely collected in several fields (e.g. Hagenaars and McCutcheon, 2002). In these settings, it is of key interest to learn dependence structures in the observed data, and to identify underlying subpopulations which may explain these patterns of dependence and their changes with external covariates. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ij})^\intercal$ denote the multivariate categorical random variable generating the observed data $\mathbf{y}_i = (y_{i1}, \ldots, y_{ij})^\intercal \in \mathcal{Y} = \{1, \ldots, K_1\} \times \cdots \times \{1, \ldots, K_J\}$, for every $i = 1, \ldots, n$. Latent class models with covariates (e.g. Bandeen-Roche et al., 1997; Formann, 1992) address this goal by assuming the responses $Y_{i1}, \ldots, Y_{ij}$ as conditionally independent given a latent class indicator $s_i \in \mathcal{S} = \{1, \ldots, R\}$, whose probability mass function is allowed to change with the covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})^\intercal \in \mathcal{X}$, under a multinomial logistic regression. According to this assumption, the conditional probability mass function $\text{pr}(\mathbf{Y}_i = \mathbf{y} \mid \mathbf{x}_i) = \text{pr}(Y_{i1} = y_1, \ldots, Y_{ij} = y_J \mid \mathbf{x}_i)$ for the multivariate random variable $\mathbf{Y}_i$ can be expressed as

$$\text{pr}(\mathbf{Y}_i = \mathbf{y} \mid \mathbf{x}_i) = \sum_{r=1}^{R} \nu_r(\mathbf{x}_i) \prod_{j=1}^{J} \pi_{jr}(y_j) = \sum_{r=1}^{R} \frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r)}{\sum_{l=1}^{R} \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_l)} \prod_{j=1}^{J} \pi_{jr}(y_j), \quad \text{for each } i = 1, \ldots, n, \tag{1}$$

where $\nu_r(\mathbf{x}_i) = \text{pr}(s_i = r \mid \mathbf{x}_i) \in (0, 1)$ is the covariate-dependent probability of class r, whereas $\pi_{jr}(y_j) = \text{pr}(Y_{ij} = y_j \mid s_i = r) \in (0, 1)$ denotes the probability to observe the category y_j for the variable Y_{ij} in class r. Consistent with classical multinomial logistic regression, the coefficients $\boldsymbol{\beta}_R = (\beta_{1R}, \ldots, \beta_{PR})^\intercal$ associated with the last class are fixed to zero, in order to avoid identifiability issues. Hence, $\mathbf{x}_i^\intercal \boldsymbol{\beta}_r$ measures the log-odds of belonging to class r instead of R, when the covariates are $\mathbf{x}_i$. Eq. (1) provides an interpretable factorization which allows inference on class-specific generative mechanisms underlying data $\mathbf{y}_i$ and how the latent classes s_i relate to the covariates $\mathbf{x}_i$. Refer to Bandeen-Roche et al. (1997) and Formann (1992) for a discussion on these models, and to Lazersfeld and Henry (1968) for an overview on latent class analysis.

* Corresponding author.
E-mail addresses: daniele.durante@unibocconi.it (D. Durante), canale@stat.unipd.it (A. Canale), tommaso.rigon@phd.unibocconi.it (T. Rigon).

https://doi.org/10.1016/j.spl.2018.10.015
0167-7152/

To obtain the above insights, it is necessary to estimate $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{R-1})$ and $\boldsymbol{\pi} = \{\pi_{j1}(y_j), \ldots, \pi_{jR}(y_j) : j = 1, \ldots, J;\ y_j = 1, \ldots, K_j\}$. This goal can be accomplished by maximizing the log-likelihood function

$$\ell(\boldsymbol{\beta}, \boldsymbol{\pi}; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} \log \left[\sum_{r=1}^{R} \nu_r(\mathbf{x}_i) \prod_{j=1}^{J} \prod_{y_j=1}^{K_j} \pi_{jr}(y_j)^{1(y_{ij}=y_j)} \right], \quad \text{with } \nu_r(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r)}{\sum_{l=1}^{R} \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_l)},\ r = 1, \ldots, R, \tag{2}$$

where $1(y_{ij} = y_j)$ is 1 if $y_{ij} = y_j$, and 0 otherwise. However, maximization of (2) is not straightforward. In fact, although early contributions attempt direct maximization of (2) via Newton–Raphson methods, more popular implementations (e.g. Bandeen-Roche et al., 1997; Formann, 1992; Van der Heijden et al., 1996; Bolck et al., 2004; Vermunt, 2010) rely on EM routines (Dempster et al., 1977). These strategies leverage a hierarchical representation, equivalent to (1), which introduces a latent variable s_i for each unit i, to obtain

$$(Y_{ij} \mid s_i = r) \sim \text{CATEGORICAL}(\boldsymbol{\pi}_{jr}, K_j), \text{ for any } j = 1, \ldots, J, \qquad (s_i \mid \mathbf{x}_i) \sim \text{CATEGORICAL}(\boldsymbol{\nu}(\mathbf{x}_i), R), \tag{3}$$

independently for every $i = 1, \ldots, n$. In (3), $\text{CATEGORICAL}(\boldsymbol{\rho}, H)$ denotes a generic categorical variable having probability mass function $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_H)$ for the H different categories. Hence, if $\mathbf{s} = (s_1, \ldots, s_n)$ is known, the maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ can be easily obtained by maximizing separately the log-likelihood $\ell_1(\boldsymbol{\beta}; \mathbf{s}, \mathbf{x})$ associated with the multinomial logistic regression for $\mathbf{s}$, and the log-likelihood $\ell_2(\boldsymbol{\pi}; \mathbf{y}, \mathbf{s})$ induced by the categorical data $\mathbf{y}_i$, $i = 1, \ldots, n$, within each class-specific subpopulation. In fact

$$\ell(\boldsymbol{\beta}, \boldsymbol{\pi}; \mathbf{y}, \mathbf{s}, \mathbf{x}) = \underbrace{\sum_{i=1}^{n} \sum_{r=1}^{R} 1(s_i = r) \log \left[\frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r)}{\sum_{l=1}^{R} \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_l)} \right]}_{\ell_1(\boldsymbol{\beta}; \mathbf{s}, \mathbf{x})} + \underbrace{\sum_{i=1}^{n} \sum_{r=1}^{R} \left[\sum_{j=1}^{J} \sum_{y_j=1}^{K_j} 1(s_i = r) 1(y_{ij} = y_j) \log \pi_{jr}(y_j) \right]}_{\ell_2(\boldsymbol{\pi}; \mathbf{y}, \mathbf{s})}. \tag{4}$$

Maximizing $\ell_1(\boldsymbol{\beta}; \mathbf{s}, \mathbf{x})$ with respect to $\boldsymbol{\beta}$ requires algorithms for multinomial logistic regression (e.g. Agresti, 2003, Chapter 7), whereas $\ell_2(\boldsymbol{\pi}; \mathbf{y}, \mathbf{s})$ is analytically maximized at

$$\hat{\pi}_{jr}(y_j) = \frac{\sum_{i=1}^{n} 1(s_i = r) 1(y_{ij} = y_j)}{\sum_{i=1}^{n} 1(s_i = r),} \quad \text{for each } r = 1, \ldots, R,\ j = 1, \ldots, J,\ y_j = 1, \ldots, K_j. \tag{5}$$

However, $\mathbf{s}$ is not observed. Thus, estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ needs to rely only on the information provided by the data. There are two main strategies to accomplish this goal, generally referred to as one-step (Bandeen-Roche et al., 1997; Formann, 1992; Van der Heijden et al., 1996) and three-step (Bolck et al., 2004; Vermunt, 2010) methods. The former attempt direct estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ exploiting (4). The latter consider, instead, a multi-step routine which first estimates $\boldsymbol{\pi}$ and $\mathbf{s}$ from a latent class model without covariates, and then uses the predicted classes $\hat{\mathbf{s}}$ as responses in $\ell_1(\boldsymbol{\beta}; \hat{\mathbf{s}}, \mathbf{x})$ – or a modification of it – to estimate $\boldsymbol{\beta}$.

Although the above methods are considered in routine implementations – including the R library `poLCA` (Linzer and Lewis, 2011) and the software LATENT GOLD (Vermunt and Magidson, 2016) – as discussed in Sections 1.1–1.2, both strategies still raise concerns on the efficiency of the algorithms and hence on the quality of the estimates. Motivated by these issues, Section 2 describes a nested EM which ensures reliable estimation for this class of models within the maximum likelihood framework. As outlined on a real dataset in Section 3, the proposed methods enjoy improved properties and performance. Section 4 provides concluding remarks.

1.1. One-step estimation methods

Recalling Section 1, maximum likelihood estimation for the parameters in the full model (1) – characterizing one-step methods – proceeds via an EM algorithm which leverages the complete log-likelihood (4). Indeed, the additive structure of (4) allows separate estimation for $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$. Moreover, $\ell_1(\boldsymbol{\beta}; \mathbf{s}, \mathbf{x})$ and $\ell_2(\boldsymbol{\pi}; \mathbf{y}, \mathbf{s})$ are linear in the augmented data $1(s_i = r)$. Letting $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\pi})$, this property facilitates a simple expectation step in which, at the general iteration t, each $1(s_i = r)$ is replaced with its conditional expectation

$$\bar{s}_{ir}^{(t)} = \frac{\exp\{\mathbf{x}_i^\intercal \boldsymbol{\beta}_r^{(t)}\} \prod_{j=1}^{J} \prod_{y_j=1}^{K_j} \pi_{jr}^{(t)}(y_j)^{1(y_{ij}=y_j)}}{\sum_{l=1}^{R} \exp\{\mathbf{x}_i^\intercal \boldsymbol{\beta}_l^{(t)}\} \prod_{j=1}^{J} \prod_{y_j=1}^{K_j} \pi_{jl}^{(t)}(y_j)^{1(y_{ij}=y_j)}}, \quad \text{for each } r = 1, \ldots, R,\ i = 1, \ldots, n, \tag{6}$$

to obtain the expected values $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$ and $Q_2(\boldsymbol{\pi} \mid \boldsymbol{\theta}^{(t)})$ of $\ell_1(\boldsymbol{\beta}; \mathbf{s}, \mathbf{x})$ and $\ell_2(\boldsymbol{\pi}; \mathbf{y}, \mathbf{s})$, respectively, whose sum defines the expectation $Q(\boldsymbol{\beta}, \boldsymbol{\pi} \mid \boldsymbol{\theta}^{(t)})$ of (4) with respect to the distribution of $\mathbf{s}$, given the current $\boldsymbol{\theta}^{(t)}$. Hence, $\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\beta}, \boldsymbol{\pi}} \{Q(\boldsymbol{\beta}, \boldsymbol{\pi} \mid \boldsymbol{\theta}^{(t)})\}$ can be obtained maximizing $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$ and $Q_2(\boldsymbol{\pi} \mid \boldsymbol{\theta}^{(t)})$ separately.

Consistent with (5), the expected log-likelihood $Q_2(\boldsymbol{\pi} \mid \boldsymbol{\theta}^{(t)})$ is easily maximized at

$$\pi_{jr}^{(t+1)}(y_j) = \frac{\sum_{i=1}^{n} \bar{s}_{ir}^{(t)} 1(y_{ij} = y_j)}{\sum_{i=1}^{n} \bar{s}_{ir}^{(t)}}, \quad \text{for each } r = 1, \ldots, R,\ j = 1, \ldots, J,\ y_j = 1, \ldots, K_j. \tag{7}$$

It is instead not possible to maximize analytically $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\beta}$, due to the logistic link. To address this issue, Formann (1992) and Van der Heijden et al. (1996) consider one Newton–Raphson step relying on a quadratic approximation of $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$. The routine proposed by Bandeen-Roche et al. (1997) leverages instead a different update relying on the Hessian and the gradient of the full-model log-likelihood (2), evaluated at $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)})$ and $\bar{s}_{ir}^{(t)}$, $r = 1, \ldots, R$, $i = 1, \ldots, n$. This procedure – currently implemented in the R library `poLCA` – breaks the EM rationale, since the M-step for $\boldsymbol{\beta}$ maximizes a plug-in estimate of a quadratic approximation for $\ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x})$, instead of $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$. This may lead to less stable behaviors.

Although the above solutions are common approaches to obtain $\boldsymbol{\beta}^{(t+1)}$, the resulting routines guarantee neither $Q_1(\boldsymbol{\beta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \geq Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$, nor $Q_1(\boldsymbol{\beta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \geq Q_1(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\theta}^{(t)})$ (Böhning and Lindsay, 1988; Böhning, 1992), and therefore the proposed methods are neither an EM, nor a generalized EM, respectively (Dempster et al., 1977). Failure to improve $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$ may also affect the monotonicity of the sequence $\ell(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}; \mathbf{y}, \mathbf{x})$, thereby providing routines which do not meet the key properties of EM, and may not ensure reliable convergence (e.g. McLachlan and Krishnan, 2007, Chapter 1.5.5). As we will outline in Section 3, this issue is not just found in pathological scenarios, but arises also in standard applications and affects quality of maximum likelihood estimation. Although careful routines can be designed to overcome this issue, we shall emphasize that $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$ is defined on a set of latent responses whose expectation (6) changes at every iteration of the algorithm. This setting is more problematic than multinomial logit with observed responses, and is further complicated by the need to estimate $\boldsymbol{\pi}$ along with $\boldsymbol{\beta}$ in (1). Hence, unstable updating steps may lead to poor estimation. Even Monte Carlo EM (Wei and Tanner, 1990) does not address the problem, since such method is devised for intractable E-steps and not for M-steps without analytical solutions.

To mitigate the above issues, standard implementations consider multiple runs based on different EM initializations, and rely on the routine converging to the highest log-likelihood. Alternatively, internal checks can be included to control for decays. These strategies provide more reliable results, but require multiple runs which increase computational costs. A possibility to reduce decays in $\ell(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}; \mathbf{y}, \mathbf{x})$, without relying on multiple runs, is to rescale the inverse of the Hessian by a step-size $0 < \alpha \leq 1$ (McLachlan and Krishnan, 2007, Chapter 1.5.6). However, few theory on optimal choices of α is available. Motivated by similar issues, Böhning and Lindsay (1988); Böhning (1992) proposed a Minorize–Majorize (MM) algorithm (e.g. Hunter and Lange, 2004) for logistic and multinomial logit regression, which replaces the Hessian with a matrix $\boldsymbol{B}$ to obtain a quadratic function minorizing the log-likelihood at every $\boldsymbol{\beta}$ and tangent to it in $\boldsymbol{\beta}^{(t)}$. This approach provides simple updating schemes which guarantee monotone log-likelihood sequences and, although not currently implemented in latent class models with covariates, can be easily incorporated in the M-step for $\boldsymbol{\beta}$. However, as outlined in Section 3, this procedure requires more iterations than nested EM.

1.2. Three-step estimation methods

Three-step methods do not attempt direct maximization of (2), but rely, instead, on a multi-step strategy which first estimates $\boldsymbol{\pi}$ and the class probabilities via closed-form EM for latent class models without covariates, and then obtain $\hat{\boldsymbol{\beta}}$ from a multinomial logit with the previously predicted classes $\hat{\mathbf{s}}$ acting as responses. This procedure allows simple and interpretable estimation, but the estimators for $\boldsymbol{\beta}$ have systematic bias.

To address this issue, Bolck et al. (2004) and Vermunt (2010), developed two bias-correction methods relying on a modification of the multinomial log-likelihood associated with $\hat{\mathbf{s}}$. The solution proposed in Bolck et al. (2004) allows estimation via standard algorithms for multinomial logit, but can be applied only when the covariates are categorical. The more general bias-correction procedure developed by Vermunt (2010) allows, instead, continuous covariates, improved efficiency and wider applicability. However, the estimates remain still sub-optimal compared to one-step methods, since they do not maximize (2), and hence cannot be considered as maximum likelihood estimates. Indeed, when the focus is on reliable inference for the parameters in (1), it is arguably more coherent to attempt a direct maximization of the log-likelihood (2), since it guarantees unbiased, efficient and consistent estimators, under the likelihood inference theory.

2. Nested EM for one-step estimation

To address the aforementioned issues, we propose a nested EM algorithm for one-step estimation which avoids approximations of the expected log-likelihood $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$, but improves this function sequentially via a set of conditional expectation–maximizations for every vector of coefficients $\boldsymbol{\beta}_r$, given the others.

Working with conditional expected log-likelihoods is appealing in providing a set of logistic functions for which the recent Pólya-gamma data augmentation scheme (Polson et al., 2013) guarantees closed-form maximization via generalized least squares. Indeed, following Polson et al. (2013), the generic logistic likelihood $\exp(\mathbf{x}^\intercal\boldsymbol{\beta})^a\{1 + \exp(\mathbf{x}^\intercal\boldsymbol{\beta})\}^{-b}$, can be rewritten as $2^{-b}\exp\{(a - 0.5b)\mathbf{x}^\intercal\boldsymbol{\beta}\}\cosh(0.5\mathbf{x}^\intercal\boldsymbol{\beta})^{-b}$, $b \geq 0$, whereas the likelihood of a Pólya-gamma variable $\omega \sim \text{PG}(b, \mathbf{x}^\intercal\boldsymbol{\beta})$ is proportional to $\exp\{-0.5\omega(\mathbf{x}^\intercal\boldsymbol{\beta})^2\}\cosh(0.5\mathbf{x}^\intercal\boldsymbol{\beta})^b$. Combining such quantities provides an augmented likelihood $\exp\{-0.5\omega(\mathbf{x}^\intercal\boldsymbol{\beta})^2 + (a - 0.5b)\mathbf{x}^\intercal\boldsymbol{\beta}\}$ proportional to the one of a linear regression for the data $\omega^{-1}(a - 0.5b) \sim \text{N}(\mathbf{x}^\intercal\boldsymbol{\beta}, \omega^{-1})$, thus leading to simple estimation of $\boldsymbol{\beta}$. Although other data augmentations for logistic regression are available, these strategies require more complex representations (Polson et al., 2013). Instead, the Pólya-gamma scheme leads to a Gaussian likelihood with a single latent variable ω, whose expectation is available via $\text{E}(\omega) = 0.5b(\mathbf{x}^\intercal\boldsymbol{\beta})^{-1}\tanh(0.5\mathbf{x}^\intercal\boldsymbol{\beta})$. Besides providing tractable computations, the proposed procedure guarantees monotone log-likelihood sequences and is directly motivated by an exact EM for the special case with $R = 2$ classes, described below.

2.1. Exact EM algorithm for the special case with $R = 2$

Let us first focus on deriving an EM for $R = 2$ latent classes, which provides analytical maximization also for $\boldsymbol{\beta}_1$. In fact, when $R = 2$, the expected value of $\ell_1(\boldsymbol{\beta}; \mathbf{s}, \mathbf{x}) = \ell_1(\boldsymbol{\beta}_1; \mathbf{s}, \mathbf{x})$ is

$$Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)}) = Q_1(\boldsymbol{\beta}_1 \mid \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \log\left[\frac{\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)^{\bar{s}_{i1}^{(t)}}}{\{1+\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)\}^{\bar{s}_{i1}^{(t)}}} \cdot \frac{\{1+\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)\}^{\bar{s}_{i1}^{(t)}}}{1+\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)}\right] = \sum_{i=1}^{n} \log\left[\frac{\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)^{\bar{s}_{i1}^{(t)}}}{1+\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)}\right],$$

thus providing a log-likelihood function $\bar{\ell}_1(\boldsymbol{\beta}_1; \bar{\mathbf{s}}^{(t)}, \mathbf{x})$ whose expression recalls the one induced by a logistic regression. This fundamental result allows the implementation of the Pólya-gamma data augmentation. In particular, defining $b := 1$, $a := \bar{s}_{i1}^{(t)}$, $\mathbf{x}^\intercal\boldsymbol{\beta} := \mathbf{x}_i^\intercal\boldsymbol{\beta}_1$ and $\omega := \omega_{i1}$, leads to the complete log-likelihood

$$\begin{aligned}\bar{\ell}_1(\boldsymbol{\beta}_1; \bar{\mathbf{s}}^{(t)}, \mathbf{x}, \boldsymbol{\omega}) &= \sum_{i=1}^{n} \log\left[\frac{\cosh\{0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)\}}{\exp\{0.5\omega_{i1}(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)^2\}} \cdot \frac{\exp\{(\bar{s}_{i1}^{(t)}-0.5)\mathbf{x}_i^\intercal\boldsymbol{\beta}_1\}}{\cosh\{0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)\}}\right] + \text{const},\\ &= \sum_{i=1}^{n}[-0.5\omega_{i1}(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1)^2 + (\bar{s}_{i1}^{(t)}-0.5)\mathbf{x}_i^\intercal\boldsymbol{\beta}_1] + \text{const}, \end{aligned} \tag{8}$$

induced by $\bar{\mathbf{s}}^{(t)}, \mathbf{x}$, and the augmented data $\boldsymbol{\omega} = (\omega_{11}, \ldots, \omega_{n1})$. Eq. (8) is a quadratic function of $\mathbf{x}_i^\intercal\boldsymbol{\beta}_1$, and is linear in $\boldsymbol{\omega}$. This structure allows the implementation of a simple nested expectation step in which every ω_{i1} is replaced with the expected value $\mathrm{E}(\omega_{i1} \mid \boldsymbol{\beta}_1^{(t)}, \mathbf{y}_i, \mathbf{x}_i) := \bar{\omega}_{i1}^{(t)} = 0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}_1^{(t)})^{-1}\tanh(0.5\mathbf{x}_i^\intercal\boldsymbol{\beta}_1^{(t)})$ to obtain $\bar{Q}_1(\boldsymbol{\beta}_1 \mid \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} -0.5\bar{\omega}_{i1}^{(t)}(\bar{\eta}_{i1}^{(t)} - \mathbf{x}_i^\intercal\boldsymbol{\beta}_1)^2 + \text{const}$, where $\bar{\eta}_{i1}^{(t)} = (\bar{s}_{i1}^{(t)} - 0.5)/\bar{\omega}_{i1}^{(t)}$. The appealing property associated with this nested expected log-likelihood, compared to $Q_1(\boldsymbol{\beta}_1 \mid \boldsymbol{\theta}^{(t)})$, is that it allows direct maximization for $\boldsymbol{\beta}_1$. Indeed, exploiting the generalized least squares, $\bar{Q}_1(\boldsymbol{\beta}_1 \mid \boldsymbol{\theta}^{(t)})$ is maximized at

$$\boldsymbol{\beta}_1^{(t+1)} = (\mathbf{X}^\intercal\bar{\boldsymbol{\Omega}}^{(t)}\mathbf{X})^{-1}\mathbf{X}^\intercal\bar{\boldsymbol{\Omega}}^{(t)}\bar{\boldsymbol{\eta}}^{(t)}, \tag{9}$$

where $\mathbf{X}$ is the $n \times P$ matrix with rows $\mathbf{x}_i^\intercal$, whereas $\bar{\boldsymbol{\Omega}}^{(t)} = \mathrm{diag}(\bar{\omega}_{11}^{(t)}, \ldots, \bar{\omega}_{n1}^{(t)})$ and $\bar{\boldsymbol{\eta}}^{(t)} = (\bar{\eta}_{11}^{(t)}, \ldots, \bar{\eta}_{n1}^{(t)})^\intercal$. Refer to Durante and Rigon (2018) for results proving that (9) guarantees monotone convergence at optimal rate in logistic models. This provides further support for nested EM, compared to Newton–Raphson and MM.

We shall stress that, although we obtain $\bar{Q}_1(\boldsymbol{\beta}_1 \mid \boldsymbol{\theta}^{(t)})$ sequentially, this quantity coincides with the expectation of the complete log-likelihood $\ell_1(\boldsymbol{\beta}_1; \mathbf{s}, \mathbf{x}, \boldsymbol{\omega})$. Hence, since $Q_2(\boldsymbol{\pi} \mid \boldsymbol{\theta}^{(t)})$ is analytically maximized in (7), the resulting routine is an exact EM algorithm based on the complete log-likelihood $\ell(\boldsymbol{\beta}_1, \boldsymbol{\pi}; \mathbf{y}, \mathbf{s}, \mathbf{x}, \boldsymbol{\omega}) = \ell_1(\boldsymbol{\beta}_1; \mathbf{s}, \mathbf{x}, \boldsymbol{\omega}) + \ell_2(\boldsymbol{\pi}; \mathbf{y}, \mathbf{s})$. Finally, note that since $\boldsymbol{\pi}^{(t+1)}$ can be easily obtained before updating $\boldsymbol{\beta}_1$, a more efficient strategy, inspired by the multi-cycle ECM (Meng and Rubin, 1993), is to update also the expectation of $\mathbf{s}$ based on $\boldsymbol{\pi}^{(t+1)}$ instead of $\boldsymbol{\pi}^{(t)}$, before applying (9). This solution will be adopted in the case with $R > 2$.

2.2. Nested EM algorithm for the general case with $R > 2$

When $R > 2$, $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$ has a multinomial logit form, and not a logistic one. Thus, direct application of Pólya-gamma data augmentation is not possible. However, as we will outline, the conditional expected log-likelihood for every $\boldsymbol{\beta}_r$, given the others, can be rewritten as a proper logistic log-likelihood (e.g. Holmes and Held, 2006), thus allowing the Pólya-gamma data augmentation. Based on this result, we propose a nested EM which improves the expected log-likelihood for $\boldsymbol{\beta}$ via a set of conditional expectation–maximizations. In particular, for each iteration t, we consider $R^* = R - 1$ nested cycles which sequentially improve the conditional expected log-likelihood of $\boldsymbol{\beta}_r$, fixing the others $\boldsymbol{\beta}_l$, $l \neq r$ at their most recent value. Hence, let $\boldsymbol{\beta}^{(t+r/R^*)} = (\boldsymbol{\beta}_1^{(t+1)}, \ldots, \boldsymbol{\beta}_r^{(t+1)}, \boldsymbol{\beta}_{r+1}^{(t)}, \ldots, \boldsymbol{\beta}_{R^*}^{(t)})$, denote the estimates for the class-specific vectors of coefficients at cycle $r = 1, \ldots, R^*$ in iteration t, we seek a sequential updating procedure providing the chain inequalities

$$Q_1(\boldsymbol{\beta}^{(t+r/R^*)} \mid \boldsymbol{\beta}^{\{t+(r-1)/R^*\}}, \boldsymbol{\pi}^{(t+1)}) \geq Q_1(\boldsymbol{\beta}^{\{t+(r-1)/R^*\}} \mid \boldsymbol{\beta}^{\{t+(r-1)/R^*\}}, \boldsymbol{\pi}^{(t+1)}), \quad \text{for each } r = 1, \ldots, R^*. \tag{10}$$

The key difference between $\boldsymbol{\beta}^{(t+r/R^*)}$ and $\boldsymbol{\beta}^{\{t+(r-1)/R^*\}}$ in (10), is that only the coefficients in $\boldsymbol{\beta}_r$ are updated, whereas all the others are kept fixed. Hence, at every cycle r we seek to improve the expected log-likelihood produced in $r-1$, by modifying only $\boldsymbol{\beta}_r$ from its previous estimate at t to a new one at $t+1$. In this respect, such strategy partially recalls the univariate version of Newton–Raphson in Vermunt and Magidson (2016), and adapts it to blocks of parameters to obtain more efficient updates relying on the most recent estimates. Moreover, as outlined in Proposition 2.1, these sequential improvements, along with the direct maximization of $Q_2(\boldsymbol{\pi} \mid \boldsymbol{\theta}^{(t)})$, guarantee the monotonicity of $\ell(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}; \mathbf{y}, \mathbf{x})$. Note that in (10), also the expectation of the conditional log-likelihood with respect to the augmented data is sequentially updated using the estimates of the coefficients from the previous cycle, in the same spirit of the multi-cycle ECM (Meng and Rubin, 1993).

In deriving an updating procedure for $\boldsymbol{\beta}$ with property (10), we adapt results in Section 2.1, which provide simple and explicit maximization for each $\boldsymbol{\beta}_r$. Focusing on cycle r within iteration t, let $Q_1(\boldsymbol{\beta}_r \mid \boldsymbol{\theta}^{\{t+(r-1)/R^*\}})$, with $\boldsymbol{\theta}^{\{t+(r-1)/R^*\}} = (\boldsymbol{\beta}^{\{t+(r-1)/R^*\}}, \boldsymbol{\pi}^{(t+1)})$, denote the conditional expected log-likelihood, written as a function of only $\boldsymbol{\beta}_r$, with all the other

Algorithm 1: Nested EM algorithm for latent class models with covariates.

Initialize the parameters $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\beta}^{(1)}$ at iteration $t = 1$. Then, for $t = 1$ until convergence of $\ell(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}; \mathbf{y}, \mathbf{x})$

Expectation: compute $\bar{s}_{ir}^{(t)}$ as in (6).

Maximization: update the current estimate $\boldsymbol{\pi}^{(t)}$ to obtain $\boldsymbol{\pi}^{(t+1)}$ as in (7).

For $r = 1$ to $r = R^*$

Nested Expectation: compute $\bar{s}_{ir}^{\{t+(r-1)/R^*\}}$ and $\bar{\omega}_{ir}^{\{t+(r-1)/R^*\}}$ by applying (6) and (13) to $\boldsymbol{\pi}^{(t+1)}$ and the current estimates of $\boldsymbol{\beta}$ produced by cycle $r-1$.

Nested Maximization: update the current estimate $\boldsymbol{\beta}_r^{(t)}$ to obtain $\boldsymbol{\beta}_r^{(t+1)}$ as in (14).

class-specific coefficients fixed at their corresponding estimates at cycle $r-1$. According to Eqs. (4) and (6), the function $Q_1(\boldsymbol{\beta}_r \mid \boldsymbol{\theta}^{\{t+(r-1)/R^*\}})$ can be expressed as

$$\sum_{i=1}^{n} \left\{ \bar{s}_{ir}^{\{t+(r-1)/R^*\}} \log \left[\frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r)}{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r) + c_i^{\{t+(r-1)/R^*\}}} \right] + \sum_{l \neq r} \bar{s}_{il}^{\{t+(r-1)/R^*\}} \log \left[\frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_l^{\{t+(r-1)/R^*\}})}{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r) + c_i^{\{t+(r-1)/R^*\}}} \right] \right\}, \tag{11}$$

where the constants $c_i^{\{t+(r-1)/R^*\}}$ denote the sum of all the exponential quantities $\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_l)$, $l \neq r$, written as a function of the current estimates for the class-specific coefficients at cycle $r-1$, whereas the expectation of the augmented latent class indicators are calculated in (6) as a function of $\boldsymbol{\pi}^{(t+1)}$ and $\boldsymbol{\beta}^{\{t+(r-1)/R^*\}}$. Since we aim to improve $Q_1(\boldsymbol{\beta}_r^{(t)} \mid \boldsymbol{\theta}^{\{t+(r-1)/R^*\}}) = Q_1(\boldsymbol{\beta}^{\{t+(r-1)/R^*\}} \mid \boldsymbol{\theta}^{\{t+(r-1)/R^*\}})$ by updating only $\boldsymbol{\beta}_r$ under the Pólya-gamma data augmentation, let us highlight a logistic log-likelihood in (11). Indeed, holding out additive constants not depending on $\boldsymbol{\beta}_r$, and dividing both the numerator and the denominator of the arguments in the logarithmic functions by the quantities $c_i^{\{t+(r-1)/R^*\}}$, $i = 1, \ldots, n$, we easily obtain

$$Q_1(\boldsymbol{\beta}_r \mid \boldsymbol{\theta}^{\{t+(r-1)/R^*\}}) = \sum_{i=1}^{n} \log \left[\frac{\{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r - a_i^{\{t+(r-1)/R^*\}})\}^{\bar{s}_{ir}^{\{t+(r-1)/R^*\}}}}{1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r - a_i^{\{t+(r-1)/R^*\}})} \right] + \text{const}, \tag{12}$$

provided that $\bar{s}_{ir}^{\{t+(r-1)/R^*\}} + \sum_{l \neq r} \bar{s}_{il}^{\{t+(r-1)/R^*\}} = 1$ and $a_i^{\{t+(r-1)/R^*\}} = \log c_i^{\{t+(r-1)/R^*\}}$. Hence, up to constants $a_i^{\{t+(r-1)/R^*\}}$, Eq. (12) for $\boldsymbol{\beta}_r$ has the same form of $Q_1(\boldsymbol{\beta}_1 \mid \boldsymbol{\theta}^{(t)})$ in Section 2.1, thereby motivating the Pólya-gamma data augmentation at each cycle r. In particular, introducing $\omega_{ir} \sim \text{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta}_r - a_i^{\{t+(r-1)/R^*\}})$, $i = 1, \ldots, n$, we can exploit the results in Section 2.1, to show that the conditional expectation of each ω_{ir} is

$$\bar{\omega}_{ir}^{\{t+(r-1)/R^*\}} = 0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r^{(t)} - a_i^{\{t+(r-1)/R^*\}})^{-1} \tanh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r^{(t)} - a_i^{\{t+(r-1)/R^*\}})], \tag{13}$$

and that the desired increment (10) at cycle r, can be simply obtained – similarly to (9) – by setting

$$\boldsymbol{\beta}_r^{(t+1)} = (\mathbf{X}^\intercal \bar{\boldsymbol{\Omega}}^{\{t+(r-1)/R^*\}} \mathbf{X})^{-1} \mathbf{X}^\intercal \bar{\boldsymbol{\Omega}}^{\{t+(r-1)/R^*\}} \bar{\boldsymbol{\eta}}^{\{t+(r-1)/R^*\}}, \tag{14}$$

where $\bar{\boldsymbol{\Omega}}^{\{t+(r-1)/R^*\}}$ is an $n \times n$ diagonal matrix with elements $\bar{\omega}_{ir}^{\{t+(r-1)/R^*\}}$, whereas $\bar{\boldsymbol{\eta}}^{\{t+(r-1)/R^*\}}$ is an $n \times 1$ vector with entries $\bar{\eta}_{ir}^{\{t+(r-1)/R^*\}} = (\bar{s}_{ir}^{\{t+(r-1)/R^*\}} - 0.5 + \bar{\omega}_{ir}^{\{t+(r-1)/R^*\}} a_i^{\{t+(r-1)/R^*\}})/\bar{\omega}_{ir}^{\{t+(r-1)/R^*\}}$, for $i = 1, \ldots, n$.

According to Algorithm 1 and the results in Proposition 2.1, all the steps of nested EM require simple and exact expressions implying monotone convergence. The proof is reported in the supplementary material.

Proposition 2.1. *The nested* EM *in Section 2.2 implies* $\ell(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\pi}^{(t+1)}; \mathbf{y}, \mathbf{x}) \geq \ell(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}; \mathbf{y}, \mathbf{x})$ *at any* t.

2.3. Hybrid nested EM *for one-step estimation*

Before evaluating the empirical performance of nested EM, we first propose a more efficient and practical hybrid version. Indeed, as discussed in Vermunt (2010), the EM is characterized by more stable maximization, whereas Newton–Raphson guarantees fast convergence when the routine is close to the maximum. Due to this, we propose a hybrid procedure which starts with nested EM and then switches to Newton–Raphson M-steps when the increment $\ell(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\pi}^{(t+1)}; \mathbf{y}, \mathbf{x}) - \ell(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}; \mathbf{y}, \mathbf{x})$ is less than or equal to a pre-specified small $\epsilon \geq 0$. Although the inclusion of Newton–Raphson steps could still cause decays in the log-likelihood sequence, when the routine is close to the maximum the parabolic approximation is more stable.

3. Empirical study

To study the benefits of nested EM in a real-data application, we compare its computational performance with those of the routines discussed in Sections 1.1 and 1.2. In particular, one-step competitors comprise the EM with one

Table 1
Quantitative evaluation of the maximization quality and the computational efficiency of the routines discussed in Sections 1.1–2. The value $\max_{\boldsymbol{\theta}}\{\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x})\}$ coincides with the highest log-likelihood observed in the 100 runs of the different algorithms. Since most of these routines are devised to maximize $\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x})$, such choice can be safely regarded as the maximum log-likelihood.

	nrEM	nrEMQ$_1$	nrEMQ$_1$ $\alpha = 0.5$	mmEM
Number of runs with a decay in $\ell(\boldsymbol{\theta}^{(t)}; \mathbf{y}, \mathbf{x})$	78	37	12	0
Number of runs reaching a local mode	94	51	24	27
$\lvert\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}, \mathbf{x}) - \max_{\boldsymbol{\theta}}\{\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x})\}\rvert$ in local modes	830.046	1105.626	4.894	0.644
Iterations to reach $\max_{\boldsymbol{\theta}}\{\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x})\}$	146	151	161	229
Averaged computational time for each run	0.073″	0.182″	0.230″	0.283″
	3stepClassical	3stepCorrection	nestedEM	hybridEM
Number of runs with a decay in $\ell(\boldsymbol{\theta}^{(t)}; \mathbf{y}, \mathbf{x})$	NA	NA	0	0
Number of runs reaching a local mode	100	100	24	25
$\lvert\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}, \mathbf{x}) - \max_{\boldsymbol{\theta}}\{\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x})\}\rvert$ in local modes	42.224	39.104	0.644	0.644
Iterations to reach $\max_{\boldsymbol{\theta}}\{\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x})\}$	NA	NA	171	166
Averaged computational time for each run	0.229″	0.212″	0.359″	0.265″

Newton–Raphson step (nrEM) from Bandeen-Roche et al. (1997), the more formal em (nrEMQ$_1$) maximizing $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)})$ (Formann, 1992; Van der Heijden et al., 1996), and its conservative version which relies on the rescaled updating of $\boldsymbol{\beta}$, as outlined in Section 1.1, with $\alpha = 0.5$ (nrEMQ$_1$, $\alpha = 0.5$). We also adapt the mm (mmEM) proposed by Böhning and Lindsay (1988) and Böhning (1992) to compare performance with a routine having monotone log-likelihood sequences. The three-step methods considered are instead the classical strategy (3stepClassical) discussed in Section 1.2, and the bias-corrected algorithm (3stepCorrection) from Vermunt (2010). Finally, we also implement the hybridEM presented in Section 2.3, setting $\epsilon = 0.01$. Although we found the results robust to moderate changes of ϵ, it is important to notice that high ϵ should be avoided since it favors Newton–Raphson steps even when $\ell(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}; \mathbf{y}, \mathbf{x})$ is far from the maximum.

To provide a detailed assessment, we perform estimation under the above algorithms for 100 different runs, with varying random initialization. The routines are all initialized at the same values – for each run – and stop when the increment in the log-likelihood sequence is less than 10^{-11}. For every run we study the maximization performance and the computational efficiency of the different algorithms. Specifically, the maximization performance is monitored by the number of runs with a decay in the log-likelihood sequence, and by the frequency of runs converging to local modes. For the runs reaching local modes, we also compute the median of the absolute difference between the log-likelihood in these local modes and the maximum one. Computational efficiency is instead studied via the median number of iterations for convergence – computed only for the runs reaching $\max_{\boldsymbol{\beta},\boldsymbol{\pi}}\{\ell(\boldsymbol{\beta}, \boldsymbol{\pi}; \mathbf{y}, \mathbf{x})\}$ – and the averaged computational time.[1] Code to reproduce the analyses and additional results are available at https://github.com/danieledurante/nEM.

Table 1 summarizes the performance of the methods discussed in Sections 1.1–2, for an application to the election data available in the R library `poLCA`. This dataset measures voters political affiliation, along with their opinions on how well six different personality traits describe the candidates Al Gore and George Bush before the 2000 presidential elections. These $J = 12$ categorical opinions are collected on a four items scale for $n = 880$ voters. Here, we assess performance of the maximization routines considering $R = 3$ classes, with the political affiliation entering as covariate in the multinomial logit for such latent classes.

Consistent with previous discussions, nestedEM and mmEM always provide monotone sequences for (2), thereby guaranteeing accurate maximization and reduced frequency of local modes. The use of Newton-type methods leads instead to decays in the log-likelihood sequences, increasing the chance of local modes far from $\max_{\boldsymbol{\beta},\boldsymbol{\pi}}\{\ell(\boldsymbol{\beta}, \boldsymbol{\pi}; \mathbf{y}, \mathbf{x})\}$. As discussed in Section 1.1, this issue is more severe for nrEM compared to the formal nrEMQ$_1$. This reduced performance of the Newton-type algorithms is mitigated by improvements in computational efficiency compared to nestedEM and mmEM, which remain, however, on a similar scale both for the averaged computational time and for the number of iterations to reach $\max_{\boldsymbol{\beta},\boldsymbol{\pi}}\{\ell(\boldsymbol{\beta}, \boldsymbol{\pi}; \mathbf{y}, \mathbf{x})\}$. Since these algorithms perform estimation in fractions of seconds, an improved maximization performance is arguably the most important property. Combining nestedEM and nrEMQ$_1$, provides an hybridEM which guarantees the accurate maximization performance of monotone algorithms and a computational efficiency comparable to Newton-type methods. Also rescaling the Newton–Raphson updating of $\boldsymbol{\beta}$ by $\alpha = 0.5$, allows improvements in maximization performance, but these gains are associated with a reduced computational efficiency. We shall also notice that the mmEM guarantees reliable maximization, but its global conservative bound requires more iterations to reach convergence compared to nestedEM and hybridEM.

As discussed in Section 1.2, the three-step algorithms do not attempt direct maximization of (2). The consequences of this are evident in Table 1 with all the runs of the 3stepClassical and the 3stepCorrection routines providing sub-optimal estimates which induce local modes in the full-model log-likelihood. Due to this, it is not possible to study the number of iterations to reach $\max_{\boldsymbol{\beta},\boldsymbol{\pi}}\{\ell(\boldsymbol{\beta}, \boldsymbol{\pi}; \mathbf{y}, \mathbf{x})\}$. Also the number of decays in $\ell(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}; \mathbf{y}, \mathbf{x})$ is somewhat irrelevant to evaluate the three-step methods, since the estimation routines are based on two separate maximizations not directly related to (2).

[1] Computations rely on R (version 3.3.2) implementations in a machine with 1 Intel Core i5 2.5 GHz processor and 4 GB ram.

It is also worth noticing that, although our studies focus on the log-likelihood sequence instead of the parameters estimates, these two quantities are directly related within a maximum likelihood approach. In fact, our goal is not on providing a new class of estimators for the parameters in (1), but an improved maximization routine for likelihood-based inference, with more reliable convergence to the optimal estimates maximizing (2). In this respect, we have also inspected the estimates associated with the local modes in Table 1, observing evident deviations from the optimal ones, especially in local maxima with a log-likelihood notably below $\max_{\boldsymbol{\beta},\boldsymbol{\pi}}\{\ell(\boldsymbol{\beta},\boldsymbol{\pi};\mathbf{y},\mathbf{x})\}$.

In performing the above studies, we initialized $\boldsymbol{\beta}$ from independent Gaussians with mean 0 and small variance 0.5. Reducing such variance – i.e. initializing $\boldsymbol{\beta}$ close to 0 – yields gains in all routines. Also in these cases, however, the performance of NESTEDEM and HYBRIDEM was not worse than the other algorithms, while being less sensitive to initialization. Similar conclusions were obtained in other applications.

4. Discussion

Motivated by the recent Pólya-gamma data augmentation for logistic regression, we developed a nested EM which ensures reliable maximum likelihood estimation of latent class models with covariates. Differently from Newton-type methods, the nested EM has theoretical guarantee of monotone log-likelihood sequences, while ensuring convergence at a faster rate than MM approaches (Durante and Rigon, 2018).

Although our focus has been on latent class analysis with covariates, the nested EM can be also adapted to multinomial logit models with Gaussian random effects. In this setting, the calculation of the expected log-likelihood for the fixed coefficients involves intractable marginalizations, thus requiring Monte Carlo EM (Wei and Tanner, 1990) or h-likelihood (Lee et al., 2006), among other methods (e.g. McCulloch, 1997). Our nested EM could improve these strategies. In fact, conditioned on Pólya-gamma augmented data, it is possible to perform closed-form marginalization and maximization as in Gaussian linear mixed models.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2018.10.015.

References

Agresti, A., 2003. Categorical Data Analysis. John Wiley & Sons.
Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., Rathouz, P.J., 1997. Latent variable regression for multiple discrete outcomes. J. Amer. Statist. Assoc. 92, 1375–1386.
Böhning, D., 1992. Multinomial logistic regression algorithm. Ann. Inst. Statist. Math. 44, 197–200.
Böhning, D., Lindsay, B.G., 1988. Monotonicity of quadratic-approximation algorithms. Ann. Inst. Statist. Math. 40, 641–663.
Bolck, A., Croon, M., Hagenaars, J., 2004. Estimating latent structure models with categorical variables: One-step versus three-step estimators. Polit. Anal. 12, 3–27.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39, 1–38.
Durante, D., Rigon, T., 2018. Conditionally conjugate mean–field variational Bayes for logistic models. ArXiv:1711.06999.
Formann, A.K., 1992. Linear logistic latent class analysis for polytomous data. J. Amer. Statist. Assoc. 87, 476–486.
Hagenaars, J.A., McCutcheon, A.L., 2002. Applied Latent Class Analysis. Cambridge University Press.
Holmes, C., Held, L., 2006. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Anal. 1, 145–168.
Hunter, D.R., Lange, K., 2004. A tutorial on MM algorithms. Am. Stat. 58, 30–37.
Lazersfeld, P.A., Henry, N.W., 1968. Latent Structure Analysis. Houghton Mifflin.
Lee, Y., Nelder, J.A., Pawitan, Y., 2006. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood. Chapman & Hall.
Linzer, D.A., Lewis, J.B., 2011. poLCA: An R package for polytomous variable latent class analysis. J. Statist. Software 42, 1–29.
McCulloch, C.E., 1997. Maximum likelihood algorithms for generalized linear mixed models. J. Amer. Statist. Assoc. 92, 162–170.
McLachlan, G., Krishnan, T., 2007. The EM Algorithm and Extensions, second ed. John Wiley & Sons.
Meng, X.L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm. Biometrika 80, 267–278.
Polson, N.G., Scott, J.G., Windle, J., 2013. Bayesian inference for logistic models using Pólya–Gamma latent variables. J. Amer. Statist. Assoc. 108, 1339–1349.
Van der Heijden, P.G., Dessens, J., Bockenholt, U., 1996. Estimating the concomitant-variable latent-class model with the EM algorithm. J. Educ. Behav. Stat. 21, 215–229.
Vermunt, J.K., 2010. Latent class modeling with covariates: Two improved three-step approaches. Polit. Anal. 18, 450–469.
Vermunt, J.K., Magidson, J., 2016. Technical Guide for Latent GOLD 5.1. Statistical Innovations Inc, Belmont, MA.
Wei, G.C., Tanner, M.A., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Amer. Statist. Assoc. 85, 699–704.