Variational Bayes estimation of hierarchical Dirichlet-multinomial mixtures for text clustering

Massimo Bilancia · Michele Di Nanni · Fabio Manca · Gianvito Pio

Abstract

In this paper, we formulate a hierarchical Bayesian version of the Mixture of Unigrams model for text clustering and approach its posterior inference through variational inference. We compute the explicit expression of the variational objective function for our hierarchical model under a mean-field approximation. We then derive the update equations of a suitable algorithm based on coordinate ascent to find local maxima of the variational target, and estimate the model parameters through the optimized variational hyperparameters. The advantages of variational algorithms over traditional Markov Chain Monte Carlo methods based on iterative posterior sampling are also discussed in detail.

Keywords Text clustering · Finite mixture models · Dirichlet-multinomial distribution · Bayesian hierarchical modelling · Variational inference

1 Introduction

Text clustering is a data analysis activity that has become increasingly important with the availability of large collections of text data from the Web (Andrews and Fox 2007; Aggarwal and Zhai 2012). Once the corpus has been suitably transformed into a structured data source, each document is assigned, in an unsupervised manner, to a single label indicating the thematic content of the document itself. The main underlying assumption is that a document can be represented as a Bag-of-Words (BOW; Harris 1954), in the sense that both the syntax and the order of occurrence of individual terms are not relevant for decoding the semantics, but only the frequency of occurrence of each term. In other words, each document can be represented as a high-dimensional vector of counts, and documents with different thematic content (topic) will tend to be dissimilar, in the sense that terms that occur frequently in documents of one class will tend to be less frequent in documents of other classes.

Since documents are converted into numerical vectors, the most obvious approach is to use classical hierarchical algorithms based on purely geometric notions, such as the distance between two points in a p-dimensional space, where clusters are sets of nested subsets arranged as a tree. Iterative partitioning algorithms, such as k-means, are another possibility. Each of these approaches has its advantages and disadvantages, both in terms of computational cost and accuracy in forming clusters, and we refer the reader to the extensive literature on this topic (Anastasiu et al. 2014; Xu and Tian 2015) for further discussion. An alternative view is the probabilistic approach, in which the overall likelihood function of a corpus is modeled as a finite mixture of Multinomial distributions, each of which is determined by a particular probability distribution of the frequency of occurrence of terms (Nigam et al. 2000). A topic is then identified as a probability distribution over the vocabulary of terms. This model, commonly known as mixture of Unigrams, has been generalized in several directions in the literature. For example, latent topic models are extremely flexible generative models that allow for multiple topics to occur simultaneously in a single document because different words in the document can be assigned to different topics (Blei et al. 2003; Blei and Lafferty 2007; Blei 2012).

Mixtures of Unigrams are finite mixtures of Multinomial likelihoods and, as such, can be represented by a hierarchical Bayesian model by introducing a set of latent variables that describe to which component of the mixture each observation belongs. The model is augmented by a set of prior distributions for the weights of the mixture and the likelihood parameters, with hyperparameters set in such a way that these distributions are weakly informative (Gelman et al. 2013). Learning the parameters of this hierarchical full-Bayes model has traditionally been dominated in the literature by the use of iterative Markov Chain Monte Carlo methods (MCMC; Frühwirth-Schnatter 2006). However, their behavior is often problematic due to the geometric properties of the likelihood surface, which is invariant for each of the $k!$ permutations of the component indices (when the mixture has k components). When the prior distributions of the model parameters

are symmetric, this invariance is transferred to the posterior distribution, with the result that any MCMC algorithm tends to jump between the posterior modes and produce inconsistent estimates. This phenomenon, known as label switching, reflects the impossibility of learning any feature of the mixture model that depends on the labels of the components (Celeux et al. 2018b). The use of MCMC methods is also problematic when we are dealing with the "big k problem" (i.e., with high-dimensional mixtures). In this case, schemes based on the Metropolis-Hastings algorithm are difficult to tune, while the use of Gibbs sampling tends to produce extremely sparse partitions when the posterior estimates of the latent variables are used to assign the observations to the components of the mixture (Chandra et al. 2020).

In contrast to MCMC methods, the variational approach to posterior inference is based on optimization. The posterior surface is approximated by a suitable non-concave objective function depending on a set of variational hyperparameters that control the quality of the approximation (Blei et al. 2017). This objective function is maximized by an ad-hoc coordinate ascent variational inference (CAVI) algorithm. Each run of the algorithm converges to a single local maximum of the objective function, and multiple runs can be used to find the best one and estimate the model parameters through the optimized variational hyperparameters. The method does not suffer from the difficulties associated with label switching, since only a single mode is explored at a time, nor is it affected by the curse of dimensionality, since the exploration of modes is limited only by the available computational resources.

Given the scenario described above, in this paper we make a number of contributions to the variational inference of generative models for textual data. First, we formulate a hierarchical Bayesian version of the Mixture of Unigrams model and approach its posterior inference through a special variational method known as mean-field inference (Plummer et al. 2020). We calculate the explicit expression of the objective function for our hierarchical model under this variational approximation. The variational target is also known as the Evidence Lower Bound (ELBO). We then derive the update equations of the CAVI algorithm to find local maxima of the ELBO. Finally, we show how these algorithmic tools can be applied in the context of Bayesian text clustering. Last but not least, we conduct a simulation experiment to investigate the goodness of the approximation of the marginal likelihood by the output of the variational procedure for the purpose of model selection (i.e., selecting the number of components of the mixture).

The paper is organized as follows. Section 2 introduces the basic notation and presents the details of our Bayesian hierarchical mixture of Unigrams. Section 3 describes the mean-field variational inference for the proposed model. Section 4 compares the advantages of optimization-based posterior variational inference with the computational difficulties of using traditional iterative MCMC algorithms. Section 5 describes the CAVI variational algorithm for posterior parameter estimation based on iterative coordinate ascent and also discusses its computational complexity. Section 6 presents some experimental results on using the proposed model for text clustering and compares it with some benchmark clustering procedures. The problem of choosing the number of components is also

investigated using a simulation study. Section 7 reports a bird's eye view of the advantages and disadvantages of variational inference, draws some conclusions, and outlines possible future work.

2 Hierarchical Dirichlet-multinomial mixtures

2.1 Basic definitions and notation

Suppose we have a vocabulary V with $p = |V|$ terms from a corpus of n documents. We make the usual hypothesis that the data generating mechanism can be viewed as a generative probabilistic model that outputs infinitely exchangeable streams of terms, such that any two finite sequences of the same length, differing only in the order of occurrence of the terms, are generated with the same probability and are considered the same BOW (Gelman et al. 2013). In practice, the BOW representation is a feature generation tool, since the i-th document, where $i = 1, 2, \ldots, n$, can be represented as a vector of counts:

$$y_i = (y_{i1}, y_{i2}, \cdots, y_{ip}), \tag{1}$$

where y_{il}, for $l = 1, 2, \ldots, p$, provides the number of occurrences for the l-th term in the vocabulary V.

In language model theory, infinite exchangeability is often reformulated in a simpler way by assuming that for any document, the probability of occurrence of a word in V does not depend on its position in the document, and that the probability of occurrence of a finite stream of words of arbitrary length can be factored as the product of the corresponding marginal probabilities. These conditions define the Unigram language model (Nigam et al. 2000), under which the likelihood of the vector of counts y_i for the i-th document takes the familiar Multinomial form:

$$p(y_i \mid \beta) = \frac{\left(\sum_{l=1}^{p} y_{il}\right)!}{\prod_{l=1}^{p} y_{il}!} \prod_{l=1}^{p} \beta_l^{y_{il}}, \tag{2}$$

where $\beta = (\beta_1, \beta_2, \cdots, \beta_p) \in \mathbb{R}^p$ is the vector of Multinomial parameters that must satisfy the constraints $\beta_l > 0$ for $l = 1, 2, \ldots, p$ and $\sum_{l=1}^{p} \beta_l = 1$. Another standard document-wise hypothesis is that the documents in the corpus are assumed to be conditionally independent given the Multinomial parameters. It is also worth pointing out the role of the Multinomial coefficient, usually abbreviated as follows:

$$\frac{\left(\sum_{l=1}^{p} y_{il}\right)!}{\prod_{l=1}^{p} y_{il}!} = \binom{y_{i+}}{y_i},$$

with $y_{i+} = y_{i1} + y_{i2} + \cdots + y_{ip}$. The parameter y_{i+} (the length of the i-th document) is a nuisance parameter that enters only into the normalization constant and is therefore irrelevant for the inference from the posterior distribution of parameters.

2.2 The hierarchical model

Extending the Unigram model to a hierarchical mixture of Unigrams introduces additional probabilistic levels to increase flexibility. We now assume that each document in the corpus can be assigned to one and only one of k different thematic contents (or topics). Under this hypothesis, the entire corpus is probabilistically modeled as a finite mixture model written as the following hierarchical specification based on the Multinomial likelihood:

$$y_i|\boldsymbol{\beta}, z_i \overset{\text{ind.}}{\sim} \text{Multinomial}_p(\boldsymbol{\beta}_j), \qquad i = 1, 2, \cdots, n, \tag{3}$$

$$z_i|\boldsymbol{\beta} \overset{\text{ind.}}{\sim} \text{Multinoulli}_k(\boldsymbol{\beta}), \qquad i = 1, 2, \cdots, n, \tag{4}$$

$$\boldsymbol{\beta}_j|\Theta \overset{\text{ind.}}{\sim} \text{Dirichlet}_p(1_p\Theta), \qquad j = 1, 2, \cdots, k, \tag{5}$$

$$\boldsymbol{\beta}|a \sim \text{Dirichlet}_k(1_k a), \tag{6}$$

where $\alpha, \theta > 0$ are strictly positive real numbers and denotes a vector of all ones. The Multinomial parameters are combined into a matrix $\boldsymbol{\beta} \in \mathbb{R}^{k\times p}$:

$$\boldsymbol{\beta} = \{\boldsymbol{\beta}_{jl}\}, \quad j = 1, 2, \ldots, k, \; l = 1, 2, \ldots, p, \tag{7}$$

where each row of the $\boldsymbol{\beta}$ matrix is a discrete probability distribution $\boldsymbol{\beta}_j \in \mathbb{R}^p$ representing a topic, i.e., a probability distribution of the vocabulary of terms V, since different documents may have different thematic content in the sense that terms that occur frequently in one document may be of little importance in another. A priori, we do not know the thematic content of each document, or in other words, we do not know what $\boldsymbol{\beta}_j$ distribution determines the probability of occurrence of words. This means that for a given document, the row index j that selects the corresponding distribution from the $\boldsymbol{\beta}$ matrix is a latent variable, which can be modeled with the latent indicator vector $z_i \in \mathbb{R}^k$ such that $z_{ij} = 1$ and $z_{ij'} = 0$ for $j' \neq j$, while $\boldsymbol{\beta} \in \mathbb{R}^k$ denotes the mixture weights (Robert 2007).

We assume that the hyperparameters Θ and are fixed and known. Since they are used to fully specify the highest level of the hierarchical structure, it is common to choose a non-informative (or otherwise weakly informative) setting, such as $\Theta = 1$. With this choice, the prior distribution (5) is uniform over the k-dimensional simplex. Of course, this is not the only possible setting. For example, Wallach et al. (2009) consider the possibility of using an asymmetric Dirichlet prior (where the hyperparameters vary across the components) and discuss the advantages of this choice. In this case, a mixed variational/empirical Bayes procedure can be used for parameter estimation. Although this situation is not the focus of this paper, we will discuss this possibility in more detail in Sect. 7. It is also worth noting that a symmetric Dirichlet prior corresponds to an exchangeable prior over the Multinomial parameters, which implicitly gives the data more weight in updating the posterior distribution of each $\boldsymbol{\beta}_j$ with an appropriate (weakly informative) choice of the

concentration parameter Θ. This model is minimal in the sense that it can provide a baseline specification to illustrate our framework for Bayesian posterior inference, a specification that can be enriched with better structured and informed priors, as discussed in more detail in Sect. 7.

The unnormalized posterior distribution of the model parameters can be factored as follows:

$$\begin{aligned} p(\beta, z, ß|y, \Theta, a) &\propto p(y|z, \beta, ß, \Theta, a)p(\beta, z, ß|\Theta, a) = \\ &= p(y|\beta, z)p(z|ß)p(\beta|\Theta)p(ß|a), \end{aligned} \tag{8}$$

where $y = (y_1, y_2, \cdots, y_n)$. The graphical representation of this dependence structure between random variables can be seen in Fig. 1. As mentioned above, the hyperparameters Θ and are fixed, but we still use them explicitly in the notation.

Taking advantage of the conditional independence between the marginal components of the likelihood and the prior distributions of the model parameters, the marginal likelihood of the model is:

$$p(y|\Theta, a) = \prod_{i=1}^{n} \int \sum_{z_i} p(y_i|\beta, z_i)p(z_i|ß)p(\beta|\Theta)p(ß|a)d ß d\beta, \tag{9}$$

which is clearly intractable. Therefore, the unnormalized posterior distribution (8) is not available in closed form, and we must resort to appropriate numerical methods for Bayesian estimation of the model parameters.

Equally interesting is the finding that the variance of each class-conditional likelihood exhibits overdispersion compared to the standard Multinomial distribution, as a consequence of explicitly accounting for the variability of the

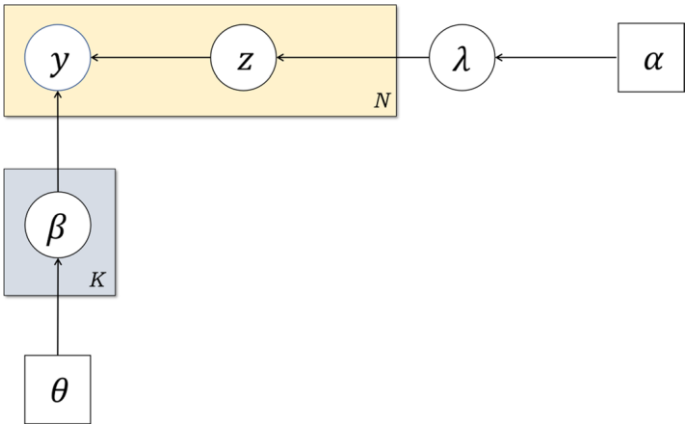

Fig. 1 Directed acyclic graph (DAG) representation of the hierarchical Dirichlet-multinomial multinomial mixture model. Circles represent stochastic nodes that may be observed (data) or unobserved (latent variables); arrows denote stochastic dependence. The number of conditionally independent components of each stochastic node (except β, which has only one component) is given in the bottom-right corner of the enclosing plates. The hyperparameters $\Theta, a \in \mathbb{R}$ enclosed in a square are fixed and known

Multinomial probabilities across topics. We give further details on this phenomenon in the Appendix 1.

3 Approximate posterior inference

To achieve our goals, we propose an optimization-based algorithm for the posterior parameter estimation of our hierarchical Dirichlet-Multinomial mixture model. To address the problem of analytic intractability of the marginal distribution (9), we will use an approach known in the literature as variational inference (Jordan et al. 1999; Blei et al. 2017). The starting point is to approximate the posterior distribution $p(\beta, z, \beta|y, \Theta, a)$ by a variational distribution $q(\beta, z, \beta|v)$, which itself depends on the variational parameters . Our optimization problem is:

$$v^{\star} = \underset{v}{\operatorname{argmin}}\, \mathbf{KL}(q(\beta, z, \lambda|v)||p(\beta, z, \lambda|y, \theta, \grave{\alpha} \tag{10}$$

where the variational objective is the reverse Kullback–Leibler (KL) divergence between the posterior distribution and the variational distribution (Murphy 2012). The solution of (10) provides the best possible approximation to the intractable posterior distribution with respect to the KL-divergence, and the optimized variational distribution $q(\beta, z, \lambda|v^{\star})$ is used to approximate the posterior inference of the model parameters. In this way, the posterior inference is treated as an optimization problem rather than a Monte Carlo sampling problem (Ghahramani 2015).

The search for the optimal approximating distribution can be greatly simplified if we define the Evidence Lower Bound (ELBO) as follows:

$$\mathrm{ELBO}(q) = \mathrm{E}_q \log p(y, z, \beta, \beta|\Theta, a) \; - \mathrm{E}_q \log q(\beta, z, \beta|v) \;, \tag{11}$$

which is a functional of the variational distribution q. By making explicit the expression of the expected values in (11), the ELBO is a function of both the variational parameters and the hyperparameters Θ and . It can then be shown that (Zhang et al. 2019; Tran et al. 2021):

$$\log p(y|\Theta, a) = \mathrm{ELBO}(q) + \mathrm{KL}(q(\beta, z, \beta|v)||p(\beta, z, \beta|y, \Theta, a)). \tag{12}$$

Since the KL-term is always positive, minimizing the KL-divergence with respect to the variational parameters is equivalent to maximizing the ELBO with respect to the variational parameters, and:

$$\log p(y|\Theta, a) \geq \mathrm{ELBO}(q),$$

i.e., the ELBO is a lower bound on the marginal log-likelihood. Therefore, minimizing (10) with respect to the variational parameters is equivalent to determining the tightest possible lower bound on the marginal log-likelihood.

3.1 Mean-field variational inference

To implement variational inference, we chose to specify the variational distribution $q(\beta, z, \beta|v)$ using a classical mean-field approximation with independent components (Wainwright and Jordan 2007; Blei et al. 2017):

$$q(\beta, z, \beta|v) = \prod_{j=1}^{k} q(\beta_j|\phi_j) \times \prod_{i=1}^{n} q(z_i|\ _i) \times q(\beta|\eta),$$

where $\phi_j \in \mathbb{R}^p$, $_i, \eta \in \mathbb{R}^k$, and:

$$\beta_j|\phi_j \overset{\text{ind.}}{\sim} \text{Dirichlet}_p(\phi_j), \qquad j = 1, 2, \cdots, k \tag{13}$$

$$z_i|\ _i \overset{\text{ind}}{\sim} \text{Multinoulli}_k(\ _i), \qquad i = 1, 2, \cdots, n \tag{14}$$

$$\beta|\eta\ \sim \text{Dirichlet}_k(\eta). \tag{15}$$

In this specification, the variational parameters $_i$ are probability distributions, while ϕ_j and η satisfy the only constraint that their components must be positive. The graphical representation of this collection of distributions, shown in Fig. 2, shows how the latent indicator variable z and the mixture weights β are decoupled in the variational distribution q (unlike in the hierarchical model formulation). Moreover, each latent indicator variable has a specific variational parameter, since we want to approximate the posterior distribution of each marginal component of the latent vector z.

For the model we are dealing with, the ELBO has the following expression (the derivation of this expression is given in Appendix 1):

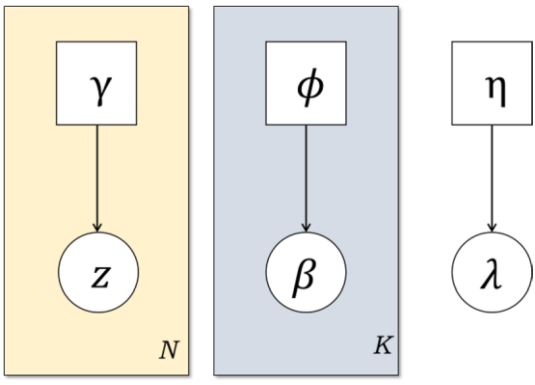

Fig. 2 Graphical representation of the mean-field variational approximation used to approximate the posterior distribution of the proposed model

$$\mathrm{ELBO}(q) = \sum_{i=1}^{n}\sum_{l=1}^{p}\sum_{j=1}^{k} y_{il}\,\phi_{ij}\left(\Psi(\alpha_{jl}) - \Psi\left(\sum_{l=1}^{p}\alpha_{jl}\right)\right) + \tag{16}$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{k} \phi_{ij}\left(\Psi(\gamma_j) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) + \tag{17}$$

$$+ k\log\Gamma(p\theta) - kp\log\Gamma(\theta) + \\ + \sum_{j=1}^{k}\sum_{l=1}^{p}(\theta-1)\left(\Psi(\alpha_{jl}) - \Psi\left(\sum_{l=1}^{p}\alpha_{jl}\right)\right)\Bigg\} + \tag{18}$$

$$+ \log\Gamma(ka) - k\log\Gamma(a) + \\ + \sum_{j=1}^{k}(a-1)\left(\Psi(\gamma_j) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)\Bigg\} - \tag{19}$$

$$- \sum_{j=1}^{k}\log\Gamma\left(\sum_{l=1}^{p}\alpha_{jl}\right) + \sum_{j=1}^{k}\sum_{l=1}^{p}\log\Gamma(\alpha_{jl}) - \\ - \sum_{j=1}^{k}\sum_{l=1}^{p}(\alpha_{jl}-1)\left(\Psi(\alpha_{jl}) - \Psi\left(\sum_{l=1}^{p}\alpha_{jl}\right)\right)\Bigg\} - \tag{20}$$

$$- \sum_{i=1}^{n}\sum_{j=1}^{k}\phi_{ij}\log\phi_{ij} - \tag{21}$$

$$- \log\Gamma\left(\sum_{j=1}^{k}\gamma_j\right) + \sum_{j=1}^{k}\log\Gamma(\gamma_j) - \\ - \sum_{j=1}^{k}(\gamma_j-1)\left(\Psi(\gamma_j) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)\Bigg\}, \tag{22}$$

where $\Psi(\cdot)$ indicates the Digamma function (logarithmic derivative of the Gamma function):

$$\Psi(z) = \frac{d}{dz}\Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}.$$

Note that the form of the class-conditional distributions is not used in computing the ELBO expression. Instead, Eqs. (16) to (19) reflect the terms of the hierarchical structure of the model, so that the computation of ELBO can be easily extended as

the model becomes more complex by adding new levels (the same holds for ELBO terms that depend on the variational distribution).

4 Sampling-based versus optimization-based posterior inference

To obtain the joint distribution of the data and latent variables, we can combine (3) and (4). Given the conditional independence assumptions, the joint distribution of data and latent indicators is:

$$p(y, z|\theta, \beta) = \prod_{i=1}^{n} p(y_i, z_i|\theta, \beta) = \prod_{i=1}^{n} p(y_i|\theta, z_i)p(z_i|\beta) = \\ = \prod_{i=1}^{n}\prod_{j=1}^{k} p(y_i|\theta_j)^{z_{ij}}\beta_j^{z_{ij}} = \prod_{i=1}^{n}\prod_{j=1}^{k} \beta_j p(y_i|\theta_j)^{z_{ij}}. \tag{23}$$

To marginalize with respect to the latent variables, we need to sum with respect to all possible z configurations, leading to the familiar likelihood representation that does not rely on the introduction of latent indicators into the model:

$$p(y_i|\theta, \beta) = \sum_{z_i^k} p(y_i|\theta, z_i)p(z_i|\beta) = \sum_{z_i} \prod_{s=1}^{k} \beta_s p(y_i|\theta_s)^{z_{is}} = \\ = \sum_{j=1}^{k} \beta_j p(y_i|\theta_j), \tag{24}$$

which is invariant for each of the $k!$ possible permutations of the summands. If we assume an exchangeable prior over the parameters of (24), this invariance is inherited by the posterior distribution, which has $k!$ symmetric modal regions corresponding to all possible permutations of the parameter labels. This phenomenon is known in the literature as 'label-switching' and causes considerable difficulty in exploring the posterior surface with MCMC sampling. Indeed, the MCMC sampler can jump between two different modes differing only in the ordering of the labels, making it impossible to compute ergodic averages to obtain posterior Monte Carlo estimates of the model parameters (Diebolt and Robert 1994). This fact reflects the impossibility of learning any feature of the distribution (24) that depends on the labels of the components.

Among the many solutions proposed in the literature, the most common is to break the exchangeability of the prior distribution by imposing constraints in the parameter space, such as $\lambda_1 < \lambda_2 < \cdots < \lambda_k$, which can be easily integrated into the MCMC sampler. However, there is no guarantee that these constraints can completely eliminate the symmetries in the posterior distribution. Other solutions work as post-processing algorithms of the MCMC output and generate a relabeling based on a suitable loss function (Celeux et al. 2000; Stephens 2000; Li and Fan 2016). It is generally accepted that these methods perform better than imposing constraints on the parameter space, since the resulting posterior marginal

distributions of the parameters are often unimodal and well separated. However, they are computationally expensive and not fully justified from a theoretical point of view, since they implicitly impose constraints that are not part of the prior specification (Kunkel and Peruggia 2020). Many alternative algorithms for relabeling have been proposed in the literature, but we do not know how they affect posterior inference and how to choose between them.

In contrast, variational inference, like any optimization algorithm, depends on initial conditions and focuses on one of the possible $k!$ modes of the posterior surface depending on those conditions. For the class of models we are concerned with, this is of course not a drawback, since we know that a single mode contains all the information for exploring latent groups and estimating parameters (Blei et al. 2017).

5 A coordinate ascent variational inference (CAVI) algorithm

A simple algorithm for maximizing the ELBO is based on a coordinate ascent scheme, where we maximize the ELBO for one parameter at a time while holding all others constant and iteratively updating the estimates until convergence is achieved (Lee 2021). For our model, it can be shown that the updating equations have a particularly simple expression. The updating equation of $_{ij}$ is, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$:

$$_{ij} \propto \exp \quad \overset{p}{\underset{l=1}{\vdots}} y_{il} \mathrm{E}_q \log \theta_{jl} \quad + \mathrm{E}_q \log \beta_j \quad . \qquad (25)$$
$$\mathcal{Z} \_\_\_\_\_ _{x_{ij}}$$

The expected values under the variational distribution q appearing in (25) coincide with the expected values of the minimal sufficient statistics of the Dirichlet distribution when represented as a natural exponential family in canonical form (Nielsen and Garcia 2009). Their explicit expression can be found in Appendix 1. Since for fixed i the set $\{ _{ij}; j = 1, 2, \ldots, k\}$ is a set of Multinational parameters that sum to 1, the proportionality sign in (25) indicates that the elements of this set must be normalized after the update is complete as follows:

$$_{ij} = \frac{\exp(x_{ij})}{\sum_{s=1}^{k} \exp(x_{is})} . \qquad (26)$$

This normalization requires special attention because the first summand in (25) is usually a very large negative number that may cause an underflow when exponentiated. In this case, a common solution is to resort to the log-sum-exp trick, which transforms the normalized values (26) to a logarithmic scale (Blanchard et al. 2021):

$$\log \phi_{ij} = \log \frac{\exp(x_{ij})}{\sum_{s=1}^{k} \exp(x_{is})} =$$

$$= \log \exp(x_{ij}) - \log \left[\sum_{s=1}^{k} \exp(x_{is}) \right] =$$

$$= x_{ij} - \log \left[\sum_{s=1}^{k} \exp(x_{is}) \right] =$$

$$= x_{ij} - \log \left[\sum_{s=1}^{k} \exp(x_{is}) \exp(Z_i) \exp(-Z_i) \right] =$$

$$= x_{ij} - Z_i - \log \left[\sum_{s=1}^{k} \exp(x_{is} - Z_i) \right] =$$

$$= \log \frac{\exp(x_{ij} - Z_i)}{\sum_{s=1}^{k} \exp(x_{is} - Z_i)},$$

and taking $Z_i = \max\{x_{i1}, x_{12}, \ldots, x_{ik}\}$ we have $\exp(x_{ij} - Z_i) \leq 1$ and:

$$\sum_{s=1}^{k} \exp(x_{is} - Z_i) \geq 1,$$

even though individual terms of the above summation may underflow to 0.

The update equations of η_j, for $j = 1, 2, \ldots, k$, and φ_{jl}, for $j = 1, 2, \ldots, k$ and $l = 1, 2, \ldots, p$, are respectively:

$$\eta_j = \alpha + \sum_{i=1}^{n} \phi_{ij}, \tag{27}$$

$$\varphi_{jl} = \theta + \sum_{i=1}^{n} y_{il} \phi_{ij}. \tag{28}$$

Proof of these update equations can also be found in Appendix 1. For the convenience of the reader, we report the pseudo-code of CAVI in Algorithm 1. In general, it can be shown that the ELBO is a concave function with respect to each of the arguments considered separately, holding all others constant (Plummer et al. 2020). Thus, the maximization of the ELBO for each parameter separately has only one solution that can be obtained with first partial derivatives without resorting to the computation of second partial derivatives or Hessian matrices. However, the ELBO is in general a non-concave function, and therefore CAVI only guarantees convergence to a local optimum that can be sensitive to the initial values (Blei et al. 2017). Results that guarantee convergence to a local maximum are known only in special cases that strongly depend on the model structure (as in the case of finite Gaussian

mixtures, Titterington and Wang 2006, or LDA, Awasthi and Risteski 2015). For the proposed model, investigating the sensitivity to the initial values by monitoring the ELBO is a reasonable diagnostic to implement variational inference with the CAVI algorithm.

Algorithm 1 CAVI algorithm for the proposed hierarchical model

Input: data $y = (y_1, y_2, \ldots, y_n)$, number of components k, prior hyperparameters $\theta, \alpha > 0$

Initialize: variational parameters γ_{ij} for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$; η_j for $j = 1, 2, \ldots, k$; $\phi_{j\ell}$ for $j = 1, 2, \ldots, k$, $\ell = 1, 2, \ldots, p$

Output: Optimized variational density $q(\beta, z, \lambda|\nu^\star) = \prod_{j=1}^{k} q(\beta_j|\phi_j^\star) \times \prod_{i=1}^{n} q(z_i|\gamma_i^\star) \times q(\lambda|\eta^\star)$

while the ELBO has not converged **do**
  for $i = 1, 2, \ldots, n$ **do**
    for $j = 1, 2 \ldots, k$ **do**
      $\gamma_{ij} = \exp\{\sum_{\ell=1}^{p} y_{i\ell} \mathsf{E}_q[\log \beta_{j\ell}]\} \exp\{\mathsf{E}_q[\log \lambda_j]\}$
      $\gamma_{ij} \leftarrow \frac{\gamma_{ij}}{\sum_{s=1}^{k} \gamma_{is}}$
    end for
  end for
  for $j = 1, 2, \ldots, k$ **do**
    $\eta_j \leftarrow \alpha + \sum_{i=1}^{n} \gamma_{ij}$
  end for
  for $j = 1, 2, \ldots, k$ **do**
    for $\ell = 1, 2, \ldots, p$ **do**
      $\phi_{j\ell} \leftarrow \theta + \sum_{i=1}^{n} y_{i\ell} \gamma_{ij}$
    end for
  end for
end while

The computational properties of the CAVI algorithm call into question the comparison with MCMC methods when the number of mixture components is very large. It is well known in the literature that MCMC algorithms have numerous structural difficulties when the number of k components is very large (Celeux et al. 2018b). In this case, the Metropolis-Hastings schemes are difficult to tune and are practically usable only for moderately sized mixtures (Frühwirth-Schnatter 2006). Gibbs sampling has been shown to be the only computationally feasible method for high-dimensional mixtures. However, several authors have noted that Gibbs sampling often fails to converge to a smaller number of nonempty clusters and promotes overfitting by providing solutions where we have many sparse clusters with only a few instances (Malsiner-Walli et al. 2016; Celeux et al. 2018b; Chandra et al. 2020). None of these problems occur with variational inference. Each run of the algorithm converges to a single local maximum of the ELBO, so we can use multiple runs to find the optimal one. Of course, this procedure is not painless, since in the case of a high-dimensional mixture we need a large number of runs, which obviously affects the total computation time. Moreover, the obtained solution may be suboptimal, since the local maxima explored may not contain the absolute maximum.

5.1 Additional considerations

The CAVI algorithm outputs the marginal components of the optimized variational distribution $q(\beta, z, \lambda|\nu^{\star})$, which can be used to obtain approximate poste- rior estimates of the model parameters. In particular, exploiting the fact that the variational distribution of each β_j is a Dirichlet:

$$\beta^{\star}_{jl} = \frac{\phi^{\star}_{jl}}{\sum_{l=1}^{p} \phi^{\star}_{jl}}, \tag{29}$$

and in the same way approximate posterior estimates of the mixing weights are given by:

$$\lambda^{\star}_{j} = \frac{\eta^{\star}_{j}}{\sum_{s=1}^{k} \eta^{\star}_{s}}. \tag{30}$$

Applying the same principle again, the approximate posterior estimate of the probability of success associated with the marginal component z_{ij} of the indicator variable z_i is calculated as follows:

$$q(z_{ij} = 1|\gamma^{\star}_{i}) = \mathrm{E}_{q}(z_{ij}|\gamma^{\star}_{i}) = \gamma^{\star}_{ij}. \tag{31}$$

The use of these probabilities becomes relevant for unsupervised classification. The decision rule is to assign the label j_i^{MAP} that satisfies the following condition:

$$j^{\mathrm{MAP}}_{i} = \underset{j}{\operatorname{argmax}}\, q(z_{ij} = 1|\gamma^{\star}_{i}). \tag{32}$$

The use of (32) to decide how to partition the sample observations is obviously justified from an intuitive point of view. Moreover, it is well known that the decision rule (32) minimizes the expected posterior loss if we use a 0/1 loss function to penalize incorrect allocations (Hastie et al. 2009).

5.2 Computational complexity

The analysis of the computational complexity of the algorithm allows us to make some quantitative considerations about its performance. In particular, looking at Algorithm 1, we can identify the following three subcomponents, for each of which we estimate the time complexity:

- $O(n * k)$ for the block 2–7 (two nested for loops), having the dominant statements at lines 4 and 5;
- $O(k)$ for the block 8–10, having the dominant statement at line 9;
- $O(k * p)$ for the block 11–14, having the dominant statement at line 13.

The time complexity of all the three subcomponents have to be multiplied by the number of iterations of the while loop, which upper bound is `maxiter` (for example, `maxiter` = 50). Since it is a constant, it is asymptotically dominated by the other terms and does not affect the overall complexity. Therefore, the whole algorithm has the following time complexity:

$$O(n * k) + O(k) + O(k * p) = O(n * k) + O(k * p). \tag{33}$$

Therefore, it is clear that if the number of documents is greater than the number of terms, $O(n * k)$ dominates the whole time complexity (and, thus, the block 2–7 is dominant); otherwise, the term $O(k * p)$ dominates the whole time complexity (and, thus, the block 11–14 is dominant). The total time complexity can also be generalized as:

$$O(\max(n, p) * k). \tag{34}$$

As explained in detail earlier, the run of the algorithm is repeated times to explore multiple modes of the ELBO surface. For the same reasons given above, the asymptotic time complexity of the algorithm does not change, since is a constant. We can also say that the running time is proportional to , i.e., it grows linearly with the number of runs. Thus, if n or p are not excessively large, Algorithm 1 remains computationally feasible even for high-dimensional mixtures, where a large number of modes will likely need to be explored to find the optimal solution.

6 Experimental work

6.1 Benchmark models

Geometric clustering algorithms Geometric clustering procedures partition the feature space into k disjoint subsets (clusters) based on the distance between any two data points. The classical algorithms used for comparisons fall into two classes (Xu and Tian 2015, is a useful survey):

- *Hierarchical agglomerative procedures*. Simple, complete and average linkage; Ward method (in each step the method finds the pair of clusters that leads to a minimum increase in total variance within the clusters after merging); Centroid method (the distance between two clusters is the distance between the two mean vectors of the clusters).
- *Iterative method*. Spherical k-means; Partitioning around medoids (PAM; unlike k-means, which uses centroids, PAM uses medoids, which are always actual points in the data set).

We measured the distance between two documents using cosine dissimilarity to eliminate the confounding effect of variability in the number of terms in each document (Dhillon and Modha 2001; Hornik et al. 2012):

$$d(y_i, y_j) = 1 - \cos(\widehat{y_i y_j}) = 1 - \frac{\langle y_i, y_j \rangle}{\|y_i\| \|y_j\|} \in [0, 2], \tag{35}$$

The Anderlucci-Viroli (AV) model More parsimonious versions of our model that can be estimated without variational inference are indeed possible. For example, Anderlucci and Viroli (2020) propose the following specification:

$$y_i | \boldsymbol{\beta}, z_i \overset{\text{ind.}}{\sim} \text{Multinomial}_p(\boldsymbol{\beta}_j), \quad i = 1, 2, \cdots, n, \tag{36}$$

$$z_i | \boldsymbol{\beta} \overset{\text{ind.}}{\sim} \text{Multinoulli}_k(\boldsymbol{\beta}), \quad i = 1, 2, \cdots, n, \tag{37}$$

$$\boldsymbol{\beta}_j | \boldsymbol{\Theta}_j \overset{\text{ind.}}{\sim} \text{Dirichlet}_p(\boldsymbol{\Theta}_j), \quad j = 1, 2, \cdots, k, \tag{38}$$

with $\boldsymbol{\Theta}_j \in \mathbb{R}^p$, where the class-conditional parameters $\boldsymbol{\Theta}_j$ and the mixture weights $\boldsymbol{\beta}$ are assumed to be fixed and unknown. For parameter estimation, the authors consider maximum likelihood estimation and propose a first-order iterative procedure based on gradient descent. The paper states that the proposed algorithm is efficient as it generally converges quickly in a few iterations. However, the dependence of the obtained solutions on the initial conditions was not further investigated, although there is a possibility that their algorithm converges to insensitive or spurious maxima or remains trapped in degeneracies of the likelihood surface (Baudry and Celeux 2015). More importantly, the proposed estimation procedure is essentially based on the fact that the class-conditional distributions are Dirichlet-Multinomials after integrating out the Multinomial parameters (see Appendix 1). We will discuss this crucial assumption in Sect. 7.

Latent Dirichlet allocation (LDA)

The specification of k probability distributions for the Multinomial parameters in the proposed model suggests an obvious similarity to the generative structure of the Latent Dirichlet Allocation (LDA) model presented in Blei et al. (2003), in which we also have k probability distributions over the vocabulary of terms V. However, with the proposed hierarchical specification, each document can be hard-clustered to a single topic. In contrast, the LDA model is a mixed membership model (Airoldi et al. 2014), and its starting point is an uncollapsed product-Multinoulli likelihood over a stream of terms, with each word associated with a latent topic. In other words, multiple topics can occur simultaneously in each document in the corpus, and a general goal of interest is to find out which themes are predominant. This model has been used extensively for the analysis of textual data (and also for the analysis of biological data; see, e.g. Sankaran and Holmes 2019, for analysis of human microbiota data based on LDA). However, it does not automatically provide better results than the standard mixture of Unigrams model and its extensions, especially for short texts or when the coexistence of multiple thematic contents is a difficult assumption to maintain.

6.2 Binary clustering of short texts

Unsupervised classification of short texts is often challenging when using traditional BOW representations due to sparse text representation (Manning et al. 2008; Rakib et al. 2020). To consider a dataset affected by these issues, we used a subset of the Reuters 21578 collection (Apté et al. 1994) previously used by Anderlucci and Viroli (2020) to compare the results of frequentist estimation of their model described in Sect. 6.1. The authors showed how their mixture of Dirichlet-Multinomials outperformed a number of standard competitors (including the Naïve Bayes mixture of Unigrams and, not surprisingly given the shortness of the texts, Latent Dirichlet Allocation). The corpus considered consists of $n = 70$ documents, 50 of which belong to the `acq` class and 20 to the `crude` class, with a clear imbalance between the two classes.

The raw text data were analyzed with R 4.2.2 (R Core Team 2022) using the infrastructure provided by the library `tm` (Feinerer et al. 2008; Feinerer and Hornik 2020). The following preprocessing steps were applied to each document in the order given: removal of extra white spaces, removal of punctuation and numbers, conversion to lowercase, removal of stop words, stemming to reduce inflectional forms to a common base form, successive recompletion using the most frequent match as completion, tokenization into unigrams (single terms). The final result is the vocabulary of terms V and the term-document matrix, whose generic element y_{il} represents the frequency of occurrence of the l-th term of V in the i-th document. Specifically, the dimension of the term-document matrix was 70×1518, with a sparsity of 96% and an average number of words per document of 54.79.

Next, we applied our model with $k = 2$ and used the default values $\alpha = 1$ and $\Theta = 5/k$. This choice is neutral with respect to and weakly informative with respect to Θ: in the absence of additional information on which to base another choice, i.e., a convincing external validation showing that a different setting has a decisive impact on the clustering process, they appear reasonable, leaving most of the responsibility for updating the posterior distribution to the data. It is also important to remember that the components of the posterior distribution are decoupled in variational inference. For example, we can change the value of without affecting the shape of the posterior distribution of β and vice versa (see also the update equations in Algorithm 1).

We applied the CAVI algorithm implemented in R 4.2.2 with `nruns`=100, starting from randomly chosen initial values, and for each of these runs we alternated between the update equations (25), (27) and (28) with `maxiter`=50. As an example, in Fig. 3 we show a subset of the trajectories of the CAVI algorithm. In general, the convergence is very fast, and it is obvious that each run reaches a different stationary point since we have many different local maxima in the ELBO. Better local optima lead to a variational approximation closer to the exact posterior. Since the ELBO is guaranteed to increase monotonically across CAVI iterations, any behavior of the trajectories that does not satisfy this requirement indicates a programming error in the code.

The results obtained were *accuracy* = 95.71% and *ARI* = 82.92 (Adjusted Rand Index), versus *accuracy* = 97.14% (*ARI* = 88.39) obtained by Anderlucci and Viroli

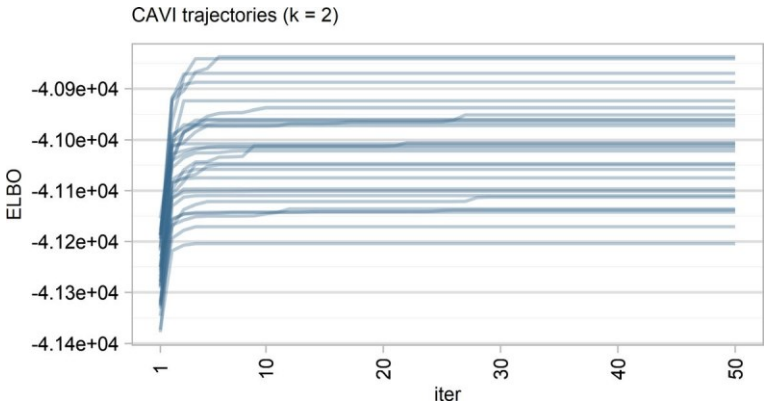

Fig. 3 Some trajectories of the CAVI algorithm applied to Reuters 21578 data ($k = 2$)

(2020). However, we repeated the numerical experiment `nruns = 500` times and obtained *accuracy* = 98.57% (*ARI* = 94.08) as the optimal result, with only one document of class `crude` incorrectly labeled as `acq`. This result suggests that there are local maxima in the ELBO surface that are not easily explored and that provide an almost exact approximation to the posterior distribution using the variational approximation. To explore these modes, we need to run the algorithm a sufficiently large number of times. However, this strategy is not computationally infeasible because, as mentioned earlier, the total computation time increases linearly as the number of runs increases.

6.3 Multiclass clustering

For this specific experiment, we adopted the dataset BBCSport, which is part of a larger collection provided for use as a benchmark for machine learning and text mining research (Greene and Cunningham 2006; Kaggle 2022). The corpus consists of $n = 737$ documents from the BBC Sport website corresponding to sports news articles in five topical areas from 2004 to 2005:

- `athletics` (101 documents, 13.70%)
- `cricket` (124 documents, 16.82%)
- `football` (265 documents, 35.96%)
- `rugby` (147 documents, 19.95%)
- `tennis` (100 documents, 13.57%)

The documents in the 5 classes show a considerable degree of semantic similarity, because although they refer to different sports, they clearly fall under the general thematic content of news published by a sports newsroom. Also in this case, the corpus was subjected to the same preprocessing steps as in the previous example, resulting in a large matrix 737×7883 with an extreme sparsity level of 98.58%. Therefore, we removed all terms that occurred less frequently, i.e., all terms $t \in V$ for which

$\mathrm{DF}_t < 0.05 \times n$, where DF_t indicates the document frequency of $t \in V$. The resulting matrix was 737×207 with an overall sparsity of 82%. Also in this case, we ran the CAVI algorithm with $k = 5$, with `nruns=100` and `maxiter=100`, $\alpha = 1$ and $\Theta = 5/k$. A subset of the trajectories is shown in Fig. 4, which confirms the impression of an extremely fast convergence to local maxima.

Unlike the previous example, where it was a simple task to assign the correct labels to the partitions created by the clustering algorithm, in this case accuracy was defined based on the best match between the true labels $c_i \in \{1, 2, \ldots, k\}$ and the cluster labels $\hat{c_i}$ as follows:

$$accuracy = \max_{p \in \mathrm{P}} \frac{1}{n} \sum_{i=1}^{n} 1\ c_i = p(\hat{c_i}), \tag{39}$$

where P is the set of all permutations in $\{1, 2, \ldots, k\}$. To solve the optimization problems in (39) in polynomial time, we used the Hungarian solver from the package `RcppHungarian` (Silverman 2022) to maximize the sum of diagonal elements of the confusion matrix with respect to all permutations of rows or columns. The results obtained are shown in Table 1, where all calculations were performed with $k = 5$ groups for comparison ($k = 5$ topics in the case of the LDA model). For the two algorithms whose results depend on the choice of initial conditions (spherical k-means and LDA), we performed 50 runs with different initial seeds, showing the best result in terms of accuracy.

The poor performance of purely geometric hierarchical methods, which cannot take into account the semantics of the problem, is of course no surprise, with some exceptions such as Ward's method, which relies on directly minimizing the total variance within each cluster, and the PAM algorithm, which iteratively assigns each document to the nearest medoid to form clusters. The best non-probabilistic method is the spherical k-means algorithm (Hornik et al. 2012), which achieves comparable performance to the analysis originally presented in Dhillon and Modha (2001). The LDA model also performs poorly, due to the aforementioned fact that each document cannot be considered a mixture of well-separated topics.

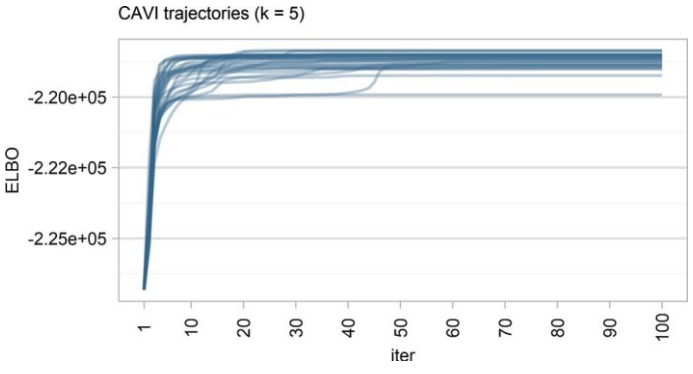

Fig. 4 Some trajectories of the CAVI algorithm applied to BBCSport data ($k = 5$)

Table 1 Adjusted Rand Index (*ARI*) and accuracy expressed as percentage (*accuracy*) for different clustering methods applied to BBCSport data

Algorithm	ARI	Accuracy (%)
Hierarchical-single linkage with cosine dissimilarity	0.00	35.96
Hierarchical-complete linkage with cosine dissimilarity	− 0.03	32.43
Hierarchical-average linkage with cosine dissimilarity	0.00	36.23
Hierarchical-Ward's method with cosine dissimilarity	0.37	58.07
Hierarchical-centroid method with cosine dissimilarity	0.00	35.96
Partitioning around medoids (PAM) with cosine dissimilarity	0.32	64.45
Spherical *k*-means with cosine dissimilarity	0.49	70.56
Latent Dirichlet allocation (LDA) with $k = 5$ topics	0.21	51.70
Hierarchical mixtures of Dirichlet-multinomials (CAVI)	0.58	74.22

Apart from the fact that the proposed model works most accurately on this dataset, it should also be noted that, unlike geometric methods, it provides estimates of $\boldsymbol{\beta}_j$ distributions (the topics) that can provide interesting clues for interpreting clusters. In Fig. 5, we provided estimates of the probability of occurrence of the top 10 terms sorting the rows of the estimated $\boldsymbol{\beta}$ by the estimated weights $\lambda_j^\star$ of each component. For example, the first distribution contains terms such as `chelsea` and `football`, which immediately point to football as the thematic content of these news (note also the estimated weight, $\lambda_2^\star = 32.71\%$ versus $\boldsymbol{\beta}_2 = 35.96\%$). The third component is clearly related to athletics, given the presence of terms such as `race` and `olympiad` ($\lambda_4^\star = 13.84\%$ versus $\boldsymbol{\beta}_4 = 13.70\%$). For the other three distributions,

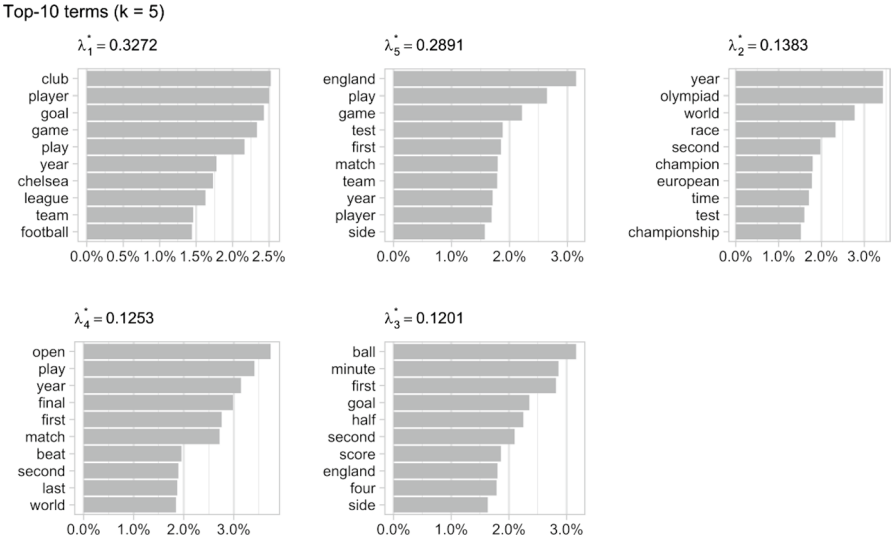

Fig. 5 Top-10 terms for each row of the estimated $\boldsymbol{\beta}$ matrix. The rows of $\boldsymbol{\beta}$ have been sorted by the estimated weights $\lambda_j^\star$ ($j = 1, 2, \ldots, k$) of each component

although the estimated weights essentially reflect the actual weights (which are not known in real applications), we have considerable difficulty in assigning the documents classified in these groups to a well-defined thematic content. This is reflected in the classification accuracy achieved (not greater than 75%), which in turn reflects the considerable overlap we can find between news about cricket and news about rugby. Tennis, of course, has its own specific terminology, but the terms that have high discriminatory power (e.g., serve, let, ace, fault) are rarely mentioned in the published news. They are often brief and generally serve only to inform about the result of the match and the statements of the participants in the post-match press release.

6.4 Model determination

In the previous examples we assumed that the number of components k was known. However, this is almost never the case, and an important question, both theoretically and practically relevant, is whether the unknown number of components k can be estimated from the output of the CAVI algorithm. The marginal likelihood is intractable in principle for our model, but could be replaced by the final value of ELBO, which is a tight lower bound for the log-marginal likelihood (as proposed, e.g., in Blei et al. 2017). However, there are some problematic issues in this context that require further discussion. First, the ELBO is certainly a lower bound on model evidence, but the variational gap varies between different models, so ELBO comparisons can be misleading. Second, for the reasons outlined in Sect. 5, it makes sense to examine only one modal region of the posterior distribution if our goal is to estimate the parameters, although this may not be sufficient if our goal is to examine the entire density surface of the marginal likelihood. This problem is known to occur with sampling-based methods, as the literature describes that traditional MCMC algorithms often do not mix appropriately and do not explore the entire support of the target distribution. Marginal likelihood estimates obtained from draws of poorly mixed samples are prone to bias (Frühwirth-Schnatter 2004; Marin and Robert 2008). These considerations also apply to variational inference, since the algorithm approximates the volume occupied by the posterior distribution of parameters by only one of the possible $k!$ modes. In Murphy (2012) a possible correction is proposed, but the extent of bias reduction is not well understood.

Another potential problem arises from the fact that the components of the approximated posterior distribution are assumed to be conditionally independent. Thus, the ELBO is an objective function that can be written as a linear sum of terms (which greatly simplifies the calculations). However, the first term (16), which is directly related to the Multinomial likelihood is very small on the logarithmic scale and dominates the other summands. Consequently, using ELBO to select k often leads to overfitting and promotes sparsity of component weights, as the optimized value of ELBO as a function of k slowly increases with increasing k, while new and poorly identified components whose mixing weights are very close to zero enter the model and increase the relative importance of the other summands. This effect is well documented in the literature dealing with

variational inference for Multinomial likelihood and is not mitigated even when we compute the predictive likelihoods on held-out data (Blei et al. 2003; Nikita 2020).

A popular alternative for model determination is the Bayesian Information Criterion (BIC; Celeux et al. 2018a), which approximates the marginal likelihood by ignoring the impact of the prior:

$$\mathsf{BIC}_k = -2l_k^\star + \mathsf{P}_k \log(n), \tag{40}$$

where $l_k^\star$ denotes the log-likelihood evaluated in the approximate parameter estimates obtained by the variational algorithm, while $\mathsf{P}_k = k(p-1) + (k-1) = kp - 1$ is the total number of free parameters of the likelihood. The BIC assumes that the data generating process is within the model collection, and it has been shown to be consistent when the probability distribution of the mixture components is bounded and satisfies mild regularity conditions (Keribin 2000). We can legitimately replace the standard maximum likelihood estimates with the variational Maximum a Posteriori (MAP) estimates and maintain the same asymptotic validity for BIC convergence. It should also be noted that, for reasons of numerical stability, it is preferable to compute the individual terms of the log-likelihood on a logarithmic scale, again by resorting to a numerical trick of the log-sum-exp type:

$$\begin{aligned} l_k^\star &= \log p(y|\beta^\star, \lambda^\star) = \log \prod_{i=1}^{n} \sum_{j=1}^{k} \lambda_j^\star p(y_i|\beta_j^\star) = \\ &= \sum_{i=1}^{n} \log \sum_{j=1}^{k} \lambda_j^\star p(y_i|\beta_j^\star) = \sum_{i=1}^{n} \log \sum_{j=1}^{k} \exp(\log(\lambda_j^\star p(y_i|\beta_j^\star))) = \\ &= \sum_{i=1}^{n} \log \sum_{j=1}^{k} \exp(\log \lambda_j^\star + \log p(y_i|\beta_j^\star)). \end{aligned} \tag{41}$$

However, given the limited precision available for standard floating-point arithmetic, it is not uncommon for the term $\log p(y_i|\beta_j^\star)$ to get into underflow, making numerical computation infeasible. In this case, the only possible solution is trade off efficiency for accuracy, by performing the computations with a library that implements floating-point arithmetic with arbitrary precision (such as Rpmfr; Maechler 2022).

Figure 6 shows an example related to the dataset BBCSport. As can be seen, the ELBO grows slowly and stabilizes only for $k \geq 9$, resulting in an overparameterization corresponding to an unparsimonious representation of the data. Conversely, the BIC strongly penalizes the number of components, leading to a slight under-parametrization with respect to the true value $k = 5$.

To further explore the model selection problem and highlight the differences between the two criteria (ELBO and BIC), we created a set of synthetic corpora obtained by subsampling the BBCSport dataset. The actual number of topics k varied in the set $\{3, 4, 5\}$, while the number of documents d in each synthetic corpus varied in the set $\{20, 50, 100\}$. For each possible pair (k, d), we created 50 corpora in the following way:

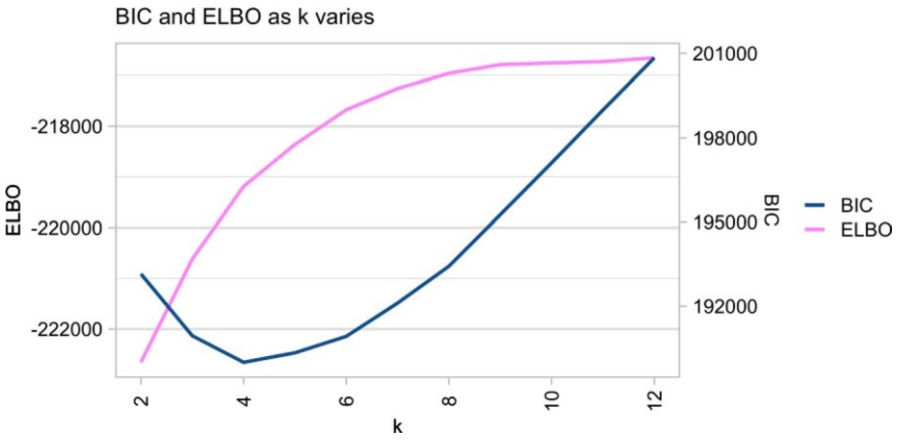

Fig. 6 Optimized ELBOs and variational estimates of the Bayesian Information Criterion (BIC) variation as a function of k (see the text for details). The dataset used is BBCSport

- To create each of the 50 corpora, we randomly extracted, without repetition, k integers between 1 and 5 (in the case $k = 5$, of course, we did no extraction, since each label appears in each corpus).
- Next, for each of the sampled labels, d documents were sampled without repetition, so that each synthetic corpus contained exactly $n = k \times d$ documents (of course, the topical areas of each of the 50 synthetic corpora may be different, except in the case $k = 5$).

For each corpus, the same preprocessing scheme was used to extract the document term matrix, closely following the scheme used for the entire BBCSport dataset. The document term matrix was created from only those terms that appeared in at least 90% of the documents. As can be easily seen in Table 2 from the relationship between p_{avg} and n, with this choice the average sparsity increases significantly with increasing k and d, so the results of the simulations are not too positively affected as the amount of information increases with increasing n.

For each (k, d) pair, we calculated the variational ELBO and the variational version of the BIC criterion for each of the 50 corpora (setting both the Dirichlet hyperparameters equal to 1; in this way the variational MAP estimate and the maximum likelihood estimates do not differ excessively). The results presented in Table 2 confirm what we have already highlighted using purely theoretical considerations. With ELBO, we obtain an overparameterization that actually increases with increasing n. For example, for $k = 5$ and $n = 500$ in 68% of the cases, the number of components is estimated to be $k^\star = 8$. Also, for $k = 3$ and $n = 60$ we have $k^\star = 8$ in 10% of the cases, a percentage that increases to 28% when $n = 300$. Conversely, the BIC criterion shows a clear tendency toward a slight under-parameterization. The extent of this under-parameterization tends to decrease as n increases. For example, at $k = 5$ we have $k^\star = 2$ in 100% of cases

Table 2 Distribution of the number of $k^\star$ components (for $k^\star$ varying between 3 and 8) estimated using either the optimized ELBO or the variational estimate of the BIC criterion. To estimate this distribution, a series of 50 synthetic corpora were drawn from BBCSports, each consisting of $n = k \times d$ documents, with $d = 20, 50, 100$ and $k = 3, 4, 5$ (see text for details)

k	n	p_{avg}	M	$k^\star = 2$ (%)	$k^\star = 3$ (%)	$k^\star = 4$ (%)	$k^\star = 5$ (%)	$k^\star = 6$ (%)	$k^\star = 7$ (%)	$k^\star = 8$ (%)
3	60	272.94	ELBO	0	16	20	22	**24**	8	10
	150	232.86		0	0	4	**28**	**28**	24	16
	300	219.66		0	0	2	6	18	**46**	28
3	60	272.94	BIC	**100**	0	0	0	0	0	0
	150	232.86		**90**	10	0	0	0	0	0
	300	219.66		22	**76**	2	0	0	0	0
4	80	249.68	ELBO	0	14	14	**28**	20	12	12
	200	221.82		0	0	0	6	30	30	**34**
	400	214.56		0	0	0	0	12	28	**60**
4	80	249.68	BIC	**100**	0	0	0	0	0	0
	200	221.82		**78**	22	0	0	0	0	0
	400	214.56		4	**50**	46	0	0	0	0
5	100	235.62	ELBO	0	4	16	12	**22**	**22**	24
	250	213.02		0	0	0	0	18	40	**42**
	500	205.98		0	0	0	0	0	32	**68**
5	100	235.62	BIC	**100**	0	0	0	0	0	0
	250	213.02		**90**	10	0	0	0	0	0
	500	205.98		0	**44**	34	22	0	0	0

k: Actual number of class labels used to resample the textual data.

n : Total number of documents in each of the 50 corpora created by subsampling the BBCSport dataset. For each label, the same number of documents was randomly selected (e.g., for $k = 4$ and $n = 80$, we have $d = 20$ documents for each topical area).

p_{avg}: average number of terms in the document-term matrices of each of the 50 resampled corpora

with $n = 100$, while at $n = 500$ the number of components is correctly set to $k^\star = 5$ in 22% of cases.

From a theoretical point of view, we know that the BIC criterion can be interpreted as a penalized estimate of the log-marginal likelihood (Murphy 2012). Also with the ELBO, the first term (16) represents an expected value of the log-likelihood (with respect to the variational distribution). The net sum of the other terms thus represents an expected penalty, which plays the same role as in the definition of the BIC criterion (41). However, this penalty is too weak compared to that of the BIC criterion, and for fixed k the importance of the first term on the logarithmic scale increases as n increases. This behavior is a consequence of the particular linear approximation of the log-marginal likelihood operated by variational inference. We will discuss these results in more detail in the next section.

7 Discussion and conclusions

Variational inference for the proposed model requires special attention, as shown in this paper. Thus, one might wonder what the real advantages of such an approach are.

The first obvious comparison is with the AV model presented in Sect. 6.1. As mentioned earlier, the estimation algorithm used for this model depends crucially on the fact that the class-conditional distribution are Dirichlet-Multinomials. In contrast, variational posterior inference is much more general, in that it does not in any way exploit the fact that the class-conditional distributions are Dirichlet-Multinomials. The ELBO is a linear sum of terms, with the Multinomial likelihood affecting only the expression of the first term. Any meaningful extension of the proposed hierarchical model is equivalent to adding new summands to the ELBO and partially reusing the existing ones. This feature is extremely important as it allows the development of new algorithms by extension and the structural complexity of the model can be easily scaled by the variational approach. Moreover, the derivation of the new update equations would not be particularly complicated, since variational inference is a first-order method that requires only partial first-order derivatives. From this point of view, we can outline several possible future research directions.

First, (5) and (6) are symmetric Dirichlet distributions with suitably specified concentration hyperparameters that give the data the responsibility of updating the posterior. In this way, unlike the classical EM algorithm, we do not have a true M-step, since the algorithm only needs to iterate along the update equations. A first obvious extension is to estimate the concentration hyperparameters Θ and directly from the data. In this case, the ELBO also depends on Θ and , and the CAVI algorithm alternately updates the variational parameters as a function of the current value of the hyperparameters and the hyperparameters as a function of the current value of the variational parameters. The expression of the ELBO we derived remains unchanged. In this way, we have a variational EM algorithm in the sense of the algorithm originally proposed by Blei et al. (2003) for the LDA model, where the final output contains both the optimized value of the ELBO

and the Empirical Bayes estimates of Θ and . Of course, in this case, the effect on the sparsity of the mixing weight distribution as the number of k components increases must be carefully considered, and in some cases the solution of estimating based on the data may not be appropriate.

Another interesting possibility is to introduce the mixing weights into a Multinomial logistic regression model ($j = 1, 2, \ldots, k$):

$$\boldsymbol{\beta}_j = \frac{\exp(u_j)}{\sum_{s=1}^{k} \exp(u_s)}, \tag{42}$$

where one of the u_j-values is set equal to zero for identifiability. The idea of structuring the weights of a mixture as in (42) first appeared in the literature in Dayton and Macready (1988) and was considered from a Bayesian perspective in Pollice and Bilancia (2000) using a generic and inefficient Gibbs sampler. Thus, the computational advantages of variational inference can be easily applied to such a model. Furthermore, if we write $u = (\tilde{u}, 0)$ and separate the zero component from the others, the values $\tilde{u}_j$ are unbounded and we can fit them into the hierarchical structure by imposing a multivariate Gaussian prior on them, as in the correlated topic model proposed in Blei and Lafferty (2007). Alternatively, we can introduce individual-specific weights $\boldsymbol{\beta}_{ij}$ as a function of a vector of concurrent variables (such as metadata) via the linear predictor $u_{ij} = \tilde{v}_j^{\top} x_i$, where $\tilde{v}_j$ is a vector of document-specific coefficients. Our model can also be easily transformed into a supervised classification algorithm following the hierarchical structure of the supervised LDA model (Zhang and Kjellström 2015).

As for the aspect of determining the number of components, the preliminary results obtained in Sect. 6.4 cast a shadow on whether ELBO can be a valid approximation to the log marginal likelihood for this purpose. Most likely, the net effect of overparameterization on the estimation is negligible, since numerous poorly identified components are introduced that have little overall weight. However, the effect on total computation time is not negligible if $k^{\star}$ becomes very high. In contrast, the BIC criterion shows a better approximation to the actual number of components, albeit with a tendency to slightly under-parameterize. Of course, these results are preliminary, and further simulations are needed, also to test other criteria showing interesting empirical performance (e.g., the Slope Heuristics, see Baudry et al. 2012). The problem of determining the number of components in a mixture in the context of variational inference remains a largely open problem. Pending further developments, a recommended empirical solution is to compare the results of the BIC criterion with those obtained by advanced dimensionality reduction methods, such as the t-distributed stochastic neighbor embedding algorithm (t-SNE; van der Maaten and Hinton 2008), which has shown particularly interesting properties in identifying the number of groups in multidimensional data.

In summary, the use of variational inference for posterior parameter estimation paves the way for a number of noteworthy developments that can be implemented with little additional effort and could greatly expand the set of tools available to

the analyst for the study of discrete multivariate data and unsupervised classification of text data.

Appendix A The Dirichlet-multinomial distribution

The j-th class-conditional distribution of the proposed hierarchical model can be written in closed form by integrating out the Multinomial parameters (in what follows $z_i = j$):

$$
\begin{aligned}
p(y_i|z_i,\Theta) &= \int p(y_i,\beta|z_i,\Theta)d\beta = \int p(y_i|\beta,z_i)p(\beta|\Theta)d\beta = \\
&= \int p(y_i|\beta_j)\prod_{s=1}^{K} p(\beta_s|\Theta)d\beta_s = \\
&= \int p(y_i|\beta_j)p(\beta_j|\Theta)d\beta_j \int p(\beta_{-j}|\Theta)d\beta_{-j} = \\
&= \binom{y_{i+}}{y_i} \prod_{l=1}^{p} \beta_{jl}^{y_{il}} \frac{\Gamma(p\Theta)}{\Gamma(\Theta)^p} \prod_{l=1}^{p} \beta_{jl}^{\Theta-1} d\beta_{jl} = \\
&= \binom{y_{i+}}{y_i} \frac{\Gamma(p\Theta)}{\Gamma(\Theta)^p} c^{-1} \int c \prod_{l=1}^{p} \beta_{jl}^{y_{il}+\Theta-1} d\beta_{jl},
\end{aligned}
$$

where the inverse of the normalization constant c has expression:

$$
c^{-1} = \frac{\prod_{l=1}^{p} \Gamma(y_{il} + \Theta)}{\Gamma\left(\sum_{l=1}^{p} y_{il} + \Theta\right)}.
$$

Using the standard notation for the multivariate Beta function:

$$
B(x) = B(x_1, x_2, \cdots, x_p) = \frac{\prod_{l=1}^{p} \Gamma(x_l)}{\Gamma\left(\sum_{l=1}^{p} x_l\right)},
$$

the class-conditional likelihood can be rewritten as:

$$
p(y_i|z_i,\Theta) = \binom{y_{i+}}{y_i} \frac{B(y_i+\Theta)}{B(\Theta)}.
$$

This probability mass function (pmf) defines the Dirichlet-Multinomial distribution. It was studied, among others, by Mosimann (1962), who showed that the variance of each marginal component of the j-th class-conditional distributions is given by:

$$\mathrm{Var}(y_{il} \mid \theta) = y_{i+}\mathrm{E}(\beta_{jl})\mathrm{E}(1 - \beta_{jl}) \frac{y_{i+} + p\Theta}{1 + p\Theta},$$

Thus, the variance of each class-conditional marginal likelihood exhibits overdispersion with respect to the standard Multinomial distribution. The magnitude of this overdispersion, which depends on the semantic heterogeneity of the underlying documents, is controlled by the term $p\Theta$, with higher values corresponding to lower overdispersion.

Appendix B Calculating the ELBO in explicit form

We begin by writing the joint distribution of the latent variables and model parameters that appears in the first term of the ELBO (11):

$$p(y, \beta, z, \beta \mid \Theta, a) = \prod_{i=1}^{n} p(y_i \mid \beta, z_i) \times \prod_{i=1}^{n} p(z_i \mid \beta) \times \prod_{j=1}^{k} p(\beta_j \mid \Theta) \times p(\beta \mid a)$$

that is:

$$\begin{aligned}
&\log p(y, \beta, z, \beta) = \\
&= \sum_{i=1}^{n} \log p(y_i \mid \beta, z_i) + \quad \boxed{\text{A1}} \\
&+ \sum_{i=1}^{n} \log p(z_i \mid \beta) + \quad \boxed{\text{A2}} \\
&+ \sum_{j=1}^{k} \log p(\beta_j \mid \Theta) + \quad \boxed{\text{A3}} \\
&+ \log p(\beta \mid a) \quad \boxed{\text{A4}}
\end{aligned}$$

We calculate the expected values of these quantities.

A1 By definition, $y_i \mid \beta, z_i \sim \mathrm{Multinomial}_p(\beta_s)$, where the index s corresponds to the index of the only component of the vector z_i that is equal to 1. It follows that:

$$\log p(y_i \mid \beta, z_i) \propto \log \prod_{l=1}^{p} \beta_{sl}^{y_{il}} = \sum_{l=1}^{p} y_{il} \log \beta_{sl},$$

and that:

$$\mathrm{E}_q \sum_{i=1}^{n} \log p(y_i|\boldsymbol{\beta}, z_i) = \mathrm{E}_q \sum_{i=1}^{n}\sum_{l=1}^{p} y_{il} \log \beta_{sl} =$$
$$= \sum_{i=1}^{n}\sum_{l=1}^{p} y_{il} \mathrm{E}_q \log \beta_{sl} =$$
$$= \sum_{i=1}^{n}\sum_{l=1}^{p}\sum_{j=1}^{k} y_{il}\, {}_{ij}\mathrm{E}_q \log \beta_{jl} ,$$

given (14), since the term $\mathrm{E}_q \log \beta_{jl}$ is a function of the random variable z_i through the index s. We now observe that the variational distribution of $\boldsymbol{\beta}_j$ can be written as:

$$q(\boldsymbol{\beta}_j|\boldsymbol{\varphi}_j) = \exp \sum_{l=1}^{p} (\varphi_{jl} - 1) \log \beta_{jl} - \sum_{l=1}^{p} \log \Gamma(\varphi_{jl}) - \log \Gamma \sum_{l=1}^{p} \varphi_{jl} ,$$

which is a multiparametric exponential family with:

- $\log \beta_{jl}$: minimal sufficient statistics for $l = 1, 2, \ldots, p$.
- $u_{jl} = \varphi_{jl} - 1$: natural (or canonical) parameters for $l = 1, 2, \ldots, p$.

By defining:

$$A(u_j) = \sum_{l=1}^{p} \log \Gamma(u_{jl} + 1) - \log \Gamma \sum_{l=1}^{p} u_{jl} + 1 ,$$

it is well known that (in what follows $\varphi_j - 1 \equiv u_j$ componentwise):

$$\mathrm{E}_q \log \beta_{jl} = \frac{\partial A(u_j)}{\partial u_{jl}} = \frac{\partial A(u_j)}{\partial \varphi_{jl}} = \frac{\partial \varphi_{jl}}{\partial u_{jl}} =$$
$$= \frac{\partial A(\varphi_j - 1)}{\partial \varphi_{jl}} \frac{\partial (u_{jl} + 1)}{\partial u_{jl}} = \frac{\partial A(\varphi_j - 1)}{\partial \varphi_{jl}} =$$
$$= \frac{\partial}{\partial \varphi_{jl}} \sum_{h=1}^{p} \log \Gamma(\varphi_{jh}) - \log \Gamma \sum_{l=1}^{p} \varphi_{jh} =$$
$$= \frac{\partial}{\partial \varphi_{jl}} \log \Gamma(\varphi_{jl}) - \frac{\partial}{\partial \varphi_{jl}} \log \Gamma \sum_{l=1}^{p} \varphi_{jh} =$$
$$= \Psi(\varphi_{jl}) - \Psi \sum_{l=1}^{p} \varphi_{jl} .$$

Putting everything together, we get the summand (16) of ELBO. □

A2 Using the independence between the latent indicator variables z_i and $\boldsymbol{\beta}$ under the variational distribution, and exploiting the representation of the variational

distribution of $\boldsymbol{\beta}$ as a multiparametric exponential family, we easily obtain the term (17):

$$\mathrm{E}_q\left[\sum_{i=1}^{n}\log p(z_i|\boldsymbol{\beta})\right] = \mathrm{E}_q\left[\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\log\beta_j\right] =$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{k}\mathrm{E}_q[z_{ij}]\,\mathrm{E}_q[\log\beta_j] =$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{k}\phi_{ij}\left(\Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k}\eta_j\right)\right).$$

□

A3 From $\boldsymbol{\beta}_j|\Theta \sim \mathrm{Dirichlet}_p(1_p\Theta)$ it readily follows that:

$$\log p(\boldsymbol{\beta}_j|\Theta) = \log\Gamma(p\Theta) - p\log\Gamma(\Theta) + \sum_{l=1}^{p}(\Theta - 1)\log\beta_{jl},$$

and:

$$\mathrm{E}_q\left[\sum_{j=1}^{k}\log p(\boldsymbol{\beta}_j|\Theta)\right] =$$
$$= \mathrm{E}_q\left[k\log\Gamma(p\Theta) - kp\log\Gamma(\Theta) + \sum_{j=1}^{k}\sum_{l=1}^{p}(\Theta - 1)\log\beta_{jl}\right] =$$
$$= k\log\Gamma(p\Theta) - kp\log\Gamma(\Theta) + \sum_{j=1}^{k}\sum_{l=1}^{p}(\Theta - 1)\,\mathrm{E}_q[\log\beta_{jl}] =$$
$$= k\log\Gamma(p\Theta) - kp\log\Gamma(\Theta) + \sum_{j=1}^{k}\sum_{l=1}^{p}(\Theta - 1)\left(\Psi(\phi_{jl}) - \Psi\left(\sum_{l=1}^{p}\phi_{jl}\right)\right),$$

that is the expression in (18). □

A4 As in the previous point, from $\boldsymbol{\beta}|\alpha \sim \mathrm{Dirichlet}_k(1_k\alpha)$ we have:

$$\log p(\boldsymbol{\beta}|\alpha) = \log\Gamma(k\alpha) - k\log\Gamma(\alpha) + \sum_{j=1}^{k}(\alpha - 1)\log\beta_j,$$

from which (19) follows that:

$$
\begin{aligned}
&\mathrm{E}_q \log p(\beta|\alpha) \;=\\
&= \log \Gamma(k\alpha) - k \log \Gamma(\alpha) + \sum_{j=1}^{k} (\alpha - 1)\mathrm{E}_q \log \beta_j \;=\\
&= \log \Gamma(k\alpha) - k \log \Gamma(\alpha) + \sum_{j=1}^{k} (\alpha - 1) \left\{ \Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k} \eta_j\right) \right\}.
\end{aligned}
$$

□

If we consider the second addend of the ELBO we have the following factorization:

$$
q(\beta, z, \beta|\gamma) = \prod_{j=1}^{k} q(\beta_j|\varphi_j) \times \prod_{i=1}^{n} q(z_i|\phi_i) \times q(\beta|\eta),
$$

that is:

$$
\begin{aligned}
&\log q(\beta, z, \beta|\gamma) =\\
&= \sum_{j=1}^{k} \log q(\beta_j|\varphi_j) + \qquad \boxed{\text{B1}}\\
&+ \sum_{i=1}^{n} \log q(z_i|\phi_i) + \qquad \boxed{\text{B2}}\\
&+ \log q(\beta|\eta) \qquad \boxed{\text{B3}}
\end{aligned}
$$

If we compute the expected value of $\log q(\beta, z, \beta|\gamma)$ with respect to the variational distribution q, using a simple algebra and the representation of the Dirichlet distribution as a multiparametric exponential family, which we have already seen, we find that the expected values with respect to q of B1, B2 and B3 correspond to (20), (21) and (22) except the sign, respectively.

Appendix C Maximizing the ELBO

Since we need to maximize each term individually, holding all others constant, we first isolate the terms in the ELBO that depend on the parameter that is being updated, and then compute the maximum point.

$\boxed{\phi_{ij}}$ ($i = 1, 2, \ldots, n, j = 1, 2, \ldots, k$). It appears in (16), (17), and (21). We isolate the factors containing ϕ_{ij} and add a Lagrangian to the objective function to account for the condition that such Multinomial parameters sum to 1 for fixed i:

$$L_{[\phi_{ij}]} = \sum_{l=1}^{p} y_{il} \left\{ \Psi(\phi_{jl}) - \Psi\left(\sum_{l=1}^{p} \phi_{jl}\right) \right\} + \\ + \left\{ \Psi(\eta_j) - \Psi\left(\sum_{l=1}^{p} \eta_j\right) \right\} - \\ - \phi_{ij} \log \phi_{ij} - L \sum_{s=1}^{k} \phi_{is} - 1 \; .$$

We take the partial derivatives to ϕ_{ij} and set them equal to zero:

$$\frac{\partial L_{[\phi_{ij}]}}{\partial \phi_{ij}} = \sum_{l=1}^{p} y_{il} \left\{ \Psi(\phi_{jl}) - \Psi\left(\sum_{l=1}^{p} \phi_{jl}\right) \right\} + \\ + \left\{ \Psi(\eta_j) - \Psi\left(\sum_{l=1}^{p} \eta_j\right) \right\} - \\ - \log \phi_{ij} - \frac{1}{\phi_{ij}} - L = 0,$$

from which we obtain:

$$\log \phi_{ij} = -1 - L + \sum_{l=1}^{p} y_{il} \left\{ \Psi(\phi_{jl}) - \Psi\left(\sum_{l=1}^{p} \phi_{jl}\right) \right\} + \\ + \left\{ \Psi(\eta_j) - \Psi\left(\sum_{l=1}^{p} \eta_j\right) \right\},$$

that is:

$$\phi_{ij} = \exp(-1 - L) \times \\ \times \exp\left\{ \sum_{l=1}^{p} y_{il} \left\{ \Psi(\phi_{jl}) - \Psi\left(\sum_{l=1}^{p} \phi_{jl}\right) \right\} \right\} \times \\ \times \exp\left\{ \Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k} \eta_j\right) \right\} \propto \\ \propto \exp\left\{ \sum_{l=1}^{p} y_{il} \mathrm{E}_q \log \beta_{jl} \right\} \exp\left\{ \mathrm{E}_q \log \beta_j \right\},$$

which must be normalized to 1 for each fixed i according to (26). □

η_j ($j = 1, 2, \ldots, k$). Isolating η_j, which appears in (17), (19) and (22), we have:

$$
\begin{aligned}
L_{[\eta_j]} = & \sum_{i=1}^{n} \gamma_{ij} \left\{ \Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k} \eta_j\right) + \right. \\
& \left. + (a-1)\left[\Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k} \eta_j\right)\right] - \right. \\
& \left. - \log \Gamma\left(\sum_{j=1}^{k} \eta_j\right) + \log \Gamma(\eta_j) - \right. \\
& \left. - (\eta_j - 1)\left[\Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k} \eta_j\right)\right] \right\} = \\
= & \sum_{i=1}^{n} \left\{ \Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k} \eta_j\right) \right\} \gamma_{ij} + a - \eta_j - \\
& - \log \Gamma\left(\sum_{j=1}^{k} \eta_j\right) + \log \Gamma(\eta_j).
\end{aligned}
$$

As above, taking the partial derivatives with respect to η_j and setting them to 0, we have:

$$
\begin{aligned}
\frac{\partial L_{[\eta_j]}}{\partial \eta_j} = & \Psi'(\eta_j) \sum_{i=1}^{n} \gamma_{ij} + a - \eta_j + \Psi(\eta_j)(-1) - \\
& - \Psi'\left(\sum_{j=1}^{k} \eta_j\right) \sum_{i=1}^{n} \gamma_{ij} + a - \eta_j - \Psi\left(\sum_{j=1}^{k} \eta_j\right)(-1) - \\
& - \Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k} \eta_j\right) = \\
= & \Psi'(\eta_j) \sum_{i=1}^{n} \gamma_{ij} + a - \eta_j - \\
& - \Psi'\left(\sum_{j=1}^{k} \eta_j\right) \sum_{i=1}^{n} \gamma_{ij} + a - \eta_j = 0,
\end{aligned}
$$

which is equivalent to the following equation in η_j:

$$
\Psi'(\eta_j) \sum_{i=1}^{n} \gamma_{ij} + a - \eta_j = \Psi'\left(\sum_{j=1}^{k} \eta_j\right) \sum_{i=1}^{n} \gamma_{ij} + a - \eta_j .
$$

For positive arguments, the Digamma function has exactly one root, so it is obvious that $\Psi'(\eta_j)$ and $\Psi'\left(\sum_{j=1}^{k} \eta_j\right)$ cannot be simultaneously zero. Therefore, this equation admits a unique solution if and only if:

$$\sum_{i=1}^{n} \quad _{ij} + a - \eta_j = 0,$$

that is if and only if:

$$\eta_j = a + \sum_{i=1}^{n} \quad _{ij}.$$

$\boxed{\varnothing_{jl}}$ ($j = 1, 2, \ldots, k$, $l = 1, 2, \ldots, p$). Isolating $\varnothing_{jl}$ in (16), (18) and (20):

$$L_{[\varnothing_{jl}]} = \sum_{i=1}^{n} y_{il} \quad _{ij} \quad \Psi(\varnothing_{jl}) - \Psi \sum_{l=1}^{p} \varnothing_{jl} \quad + \Big\}$$

$$+ (\Theta - 1) \quad \Psi(\varnothing_{jl}) - \Psi \sum_{l=1}^{p} \varnothing_{jl} \quad -$$

$$- \log \Gamma \sum_{l=1}^{p} \varnothing_{jl} \quad + \log \Gamma(\varnothing_{jl}) -$$

$$- (\varnothing_{jl} - 1) \quad \Psi(\varnothing_{jl}) - \Psi \sum_{l=1}^{p} \varnothing_{jl} \quad \Big\} =$$

$$= \quad \Psi(\varnothing_{jl}) - \Psi \sum_{l=1}^{p} \varnothing_{jl} \quad \Big\} \sum_{i=1}^{n} y_{il} \quad _{ij} + \Theta - \varnothing_{jl} \quad -$$

$$- \log \Gamma \sum_{l=1}^{p} \varnothing_{jl} \quad + \log \Gamma(\varnothing_{jl}).$$

Taking the first derivative and setting it to 0:

$$\frac{6L_{[\varnothing_{jl}]}}{6\varnothing_{jl}} = \Psi^{f}(\varnothing_{jl}) \sum_{i=1}^{n} y_{il} \quad _{ij} + \Theta - \varnothing_{jl} \quad +$$

$$+ \Psi(\varnothing_{jl})(-1) - \Psi \sum_{l=1}^{p} \varnothing_{jl} \quad \sum_{i=1}^{n} y_{il} \quad _{ij} + \Theta - \varnothing_{jl} \quad -$$

$$- \Psi^{f} \sum_{l=1}^{p} \varnothing_{jl} \quad (-1) - \Psi \sum_{l=1}^{p} \varnothing_{jl} \quad + \Psi_{jl} =$$

$$= \Psi^{f}(\varnothing_{jl}) \sum_{i=1}^{n} y_{il} \quad _{ij} + \Theta - \varnothing_{jl} \quad -$$

$$- \Psi^{f} \sum_{l=1}^{p} \varnothing_{jl} \quad \sum_{i=1}^{n} y_{il} \quad _{ij} + \Theta - \varnothing_{jl} \quad = 0,$$

which, as in the previous case, it has a unique solution in $\varnothing_{jl}$ given by:

$$\varnothing_{jl} = \Theta + \sum_{i=1}^{n} y_{il}\ _{ij}.$$

Acknowledgements We wish to thank the Associate Editor for his help and support and the two anonymous referees for their careful and constructive reviews.

Author contributions The authors contributed to the manuscript equally.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability Available on the websites referenced in the article.

Code availability Upon request.

Declarations

Conflict of interest The authors have no conflict of interest to disclose.

References

Aggarwal CC, Zhai C (2012) Mining text data. Springer, New York. https://doi.org/10.1007/978-1-4614-3223-4

Airoldi EM, Blei D, Erosheva EA et al (2014) Handbook of mixed membership models and their applications. Chapman and Hall, Boca Raton. https://doi.org/10.1201/b17520

Anastasiu DC, Tagarelli A, Karypis G (2014) Document clustering: the next frontier. In: Aggarwal CC, Reddy CK (eds) Data clustering: algorithms and applications. Chapman & Hall, Boca Raton, pp 305–338

Anderlucci L, Viroli C (2020) Mixtures of Dirichlet-multinomial distributions for supervised and unsupervised classification of short text data. Adv Data Anal Classif 14:759–770. https://doi.org/10.1007/s11634-020-00399-3

Andrews N, Fox E (2007) Recent developments in document clustering. http://hdl.handle.net/10919/19473, Virginia Tech computer science technical report, TR-07-35

Apté C, Damerau F, Weiss SM (1994) Automated learning of decision rules for text categorization. ACM Trans Inf Syst 12:233–251. https://doi.org/10.1145/183422.183423

Awasthi P, Risteski A (2015) On some provably correct cases of variational inference for topic models. In: Cortes C, Lawrence N, Lee D et al (eds) Advances in neural information processing systems, vol 28. Curran Associates, Inc., New York

Baudry JP, Celeux G (2015) EM for mixtures. Inizialiation requires special care. Stat Comput 25:713–726. https://doi.org/10.1007/s11222-015-9561-x

Baudry JP, Maugis C, Michel B (2012) Slope heuristics: overview and implementation. Stat Comput 22:455–470. https://doi.org/10.1007/s11222-011-9236-1

Blanchard P, Higham DJ, Higham NJ (2021) Accurately computing the log-sum-exp and softmax functions. IMA J Numer Anal 41:2311–2330. https://doi.org/10.1093/imanum/draa038

Blei DM (2012) Probabilistic topic models. Commun ACM 55:77–84. https://doi.org/10.1145/2133806.2133826

Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat. https://doi.org/10.1214/07-AOAS114

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: a review for statisticians. J Am Stat Assoc 112:859–877. https://doi.org/10.1080/01621459.2017.1285773

Celeux G, Hurn M, Robert CP (2000) Computational and inferential difficulties with mixture posterior distributions. J Am Stat Assoc 95:957–970. https://doi.org/10.1080/01621459.2000.10474285
Celeux G, Früwirth-Schnatter S, Robert CP (2018a) Model selection for mixture models—perspectives and strategies. In: Frühwirth-Schnatter S, Celeux G, Robert CP (eds) Handbook of mixture analysis. Chapmann & Hall, New York, pp 118–154. https://doi.org/10.1201/9780429055911
Celeux G, Kamary K, Malsiner-Walli G et al (2018b) Computational solutions for Bayesian inference in mixture models. In: Frühwirth-Schnatter S, Celeux G, Robert CP (eds) Handbook of mixture analysis. Chapmann & Hall, New York, pp 73–115. https://doi.org/10.1201/9780429055911
Chandra NK, Canale A, Dunson DB (2020) Escaping the curse of dimensionality in Bayesian model based clustering. arxiv:2006.02700
Dayton CM, Macready GB (1988) Concomitant-variable latent-class models. J Am Stat Assoc 83:173. https://doi.org/10.2307/2288938
Dhillon IS, Modha DS (2001) Concept decompositions for large sparse text data using clustering. Mach Learn 42:143–175. https://doi.org/10.1023/A:1007612920971
Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J R Stat Soc Ser B (Methodol) 56:363–375. https://doi.org/10.1111/j.2517-6161.1994.tb01985.x
Feinerer I, Hornik K, Meyer D (2008) Text mining infrastructure in R. J Stati Softw. https://doi.org/10.18637/jss.v025.i05
Feinerer I, Hornik K (2020) tm: text mining package. https://CRAN.R-project.org/package=tm, R package version 0.7-8
Frühwirth-Schnatter S (2004) Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. Econom J 7:143–167. https://doi.org/10.1111/j.1368-423X.2004.00125.x
Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, New York. https://doi.org/10.1007/978-0-387-35768-3
Gelman A, Carlin J, Stern H et al (2013) Bayesian data analysis, 3rd edn. Chapman and Hall, Boca Raton
Ghahramani Z (2015) Probabilistic machine learning and artificial intelligence. Nature 521:452–459. https://doi.org/10.1038/nature14541
Greene D, Cunningham P (2006) Practical solutions to the problem of diagonal dominance in kernel document clustering. In: Proceedings of the 23rd international conference on machine learning (ICML'06). ACM Press, pp 377–384
Harris ZS (1954) Distributional structure. WORD 10:146–162. https://doi.org/10.1080/00437956.1954.11659520
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7
Hornik K, Feinerer I, Kober M et al (2012) Spherical *k*-means clustering. J Stat Softw. https://doi.org/10.18637/jss.v050.i10
Jordan MI, Ghahramani Z, Jaakkola TS et al (1999) An introduction to variational methods for graphical models. Mach Learn 37:183–233. https://doi.org/10.1023/A:1007665907178
Kaggle (2022) Sports dataset(bbc). https://www.kaggle.com/datasets/maneesh99/sports-datasetbbc. Accessed 04 Nov 2022
Keribin C (2000) Consistent estimation of the order of mixture models. Sankhyā Indian J Stat Ser A (1961–2002) 62:49–66
Kunkel D, Peruggia M (2020) Anchored Bayesian Gaussian mixture models. Electron J Stat. https://doi.org/10.1214/20-EJS1756
Lee SY (2021) Gibbs sampler and coordinate ascent variational inference: a set-theoretical review. Commun Stat Theory Methods. https://doi.org/10.1080/03610926.2021.1921214
Li H, Fan X (2016) A pivotal allocation-based algorithm for solving the label-switching problem in Bayesian mixture models. J Comput Graph Stat 25:266–283. https://doi.org/10.1080/10618600.2014.983643
Maechler M (2022) Rmpfr: R mpfr—multiple precision floating-point reliable. https://cran.r-project.org/package=Rmpfr, R package version 0.8-9
Malsiner-Walli G, Frühwirth-Schnatter S, Grün B (2016) Model-based clustering based on sparse finite Gaussian mixtures. Stat Comput 26:303–324. https://doi.org/10.1007/s11222-014-9500-2
Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
Marin JM, Robert C (2008) Approximating the marginal likelihood in mixture models. Indian Bayesian Soc Newslett 5:2–7

Mosimann JE (1962) On the compound multinomial distribution, the multivariate Beta-distribution, and correlations among proportions. Biometrika 49:65–82. https://doi.org/10.1093/biomet/49.1-2.65
Murphy KP (2012) Machine learning: a probabilistic perspective. The MIT Press, Cambridge
Nielsen F, Garcia V (2009) Statistical exponential families: a digest with flash cards. arXiv:0911.4863
Nigam K, Mccallum AK, Thrun S et al (2000) Text classification from labeled and unlabeled documents using EM. Mach Learn 39:103–134. https://doi.org/10.1023/A:1007692713085
Nikita M (2020) ldatuning: tuning of the latent Dirichlet allocation models parameters. https://CRAN.R-project.org/package=ldatuning, R package version 1.0.2
Plummer S, Pati D, Bhattacharya A (2020) Dynamics of coordinate ascent variational inference: a case study in 2D Ising models. Entropy 22:1263. https://doi.org/10.3390/e22111263
Pollice A, Bilancia M (2000) A hierarchical finite mixture model for Bayesian classification in the presence of auxiliary information. Metron Int J Stat LVIII:109–131
R Core Team (2022) R: a language and environment for statistical computing. https://www.R-project.org/
Rakib MRH, Zeh N, Jankowska M et al (2020) Enhancement of short text clustering by iterative classification. In: Métais E, Meziane F, Horacek H et al (eds) Natural language processing and information systems. Springer, Berlin, pp 105–117. https://doi.org/10.1007/978-3-030-51310-8_10
Robert CP (2007) The Bayesian choice. Springer, New York. https://doi.org/10.1007/0-387-71599-1
Sankaran K, Holmes SP (2019) Latent variable modeling for the microbiome. Biostatistics 20:599–614. https://doi.org/10.1093/biostatistics/kxy018
Silverman J (2022) RcppHungarian: solves minimum cost bipartite matching problems. https://CRAN.R-project.org/package=RcppHungarian, R package version 0.2
Stephens M (2000) Dealing with label switching in mixture models. J R Stat Soc Ser B (Stat Methodol) 62:795–809. https://doi.org/10.1111/1467-9868.00265
Titterington DM, Wang B (2006) Convergence properties of a general algorithm for calculating variational Bayesian estimates for a Normal mixture model. Bayesian Anal. https://doi.org/10.1214/06-BA121
Tran MN, Nguyen TN, Dao VH (2021) A practical tutorial on variational Bayes. arXiv:2103.01327
van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
Wainwright MJ, Jordan MI (2007) Graphical models, exponential families, and variational inference. Found Trends® Mach Learn 1:1–305. https://doi.org/10.1561/2200000001
Wallach H, Mimno D, McCallum A (2009) Rethinking LDA: why priors matter. In: Bengio Y, Schuurmans D, Lafferty J et al (eds) Advances in neural information processing systems, vol 22. Curran Associates Inc., New York
Xu D, Tian Y (2015) A comprehensive survey of clustering algorithms. Ann Data Sci 2:165–193. https://doi.org/10.1007/s40745-015-0040-1
Zhang C, Butepage J, Kjellstrom H et al (2019) Advances in variational inference. IEEE Trans Pattern Anal Mach Intell 41:2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774
Zhang C, Kjellström H (2015) How to supervise topic models. In: Agapito L, Bronstein MM, Rother C (eds) Computer vision—ECCv 2014 workshops. Springer, Cham, pp 500–515. https://doi.org/10.1007/978-3-319-16181-5_39