Statistical Papers (2025) 66:80
https://doi.org/10.1007/s00362-025-01702-0

REGULAR ARTICLE

Stochastic variational inference for clustering short text data with finite mixtures of Dirichlet-Multinomial distributions

Massimo Bilancia[1] · **Andrea Nigri**[2] · **Samuele Magro**[3]

Received: 1 August 2024 / Revised: 25 February 2025

Abstract

Finite mixtures of Multinomial distributions are a valuable tool for analyzing discrete positive data, particularly in the context of text analysis where data is represented as a Bag-of-Words (BOW). In this approach, only term frequency from a predefined vocabulary is considered, disregarding the specific positions of terms within the pre-processed text document. Dirichlet-Multinomial mixture models, in particular, offer a straightforward yet effective method for text categorization. These models often outperform more complex latent variable models in cases where documents are short. The combination of Dirichlet priors and Multinomial likelihoods can be addressed within a Bayesian framework. However, despite the model's simplicity, the exact posterior distribution is intractable, necessitating the use of numerical methods. Variational inference offers a promising approach by approximating the joint posterior distribution with a probability distribution in which the model parameters are assumed to be independent a posteriori. Under certain conditions, a coordinate ascent variational algorithm can be constructed to yield an approximation that closely matches the true posterior in terms of the reverse Kullback–Leibler divergence. A notable limitation of standard variational algorithms, however, is their requirement to use the entire dataset to compute the iterative equations for estimating the local variational parameters, which poses a significant scalability issue when working with large text corpora. To address this, we employ stochastic variational inference within the exponential family

✉ Massimo Bilancia
massimo.bilancia@uniba.it

Andrea Nigri
andrea.nigri@unifg.it

Samuele Magro
magrosamuele1999@gmail.com

[1] Department of Precision and Regenerative Medicine and Jonian Area (DiMePRe-J), University of Bari Aldo Moro, Polyclinic University Hospital, Piazza Giulio Cesare 11, 70124 Bari, Italy

[2] Department of Economics, Management and Territory (DEMeT), University of Foggia, Via Alberto da Zara 11, 71121 Foggia, Italy

[3] Sirio Astronomical Observatory, Grotte di Castellana srl, Piazzale Anelli, 70013 Castellana Grotte, Italy

Published online: 25 April 2025

to develop a scalable estimation algorithm. By leveraging straightforward assumptions about the full conditional distributions of the hierarchical model and the distributions of the variational parameters, we demonstrate that, under the Robbins–Monro conditions, a gradient ascent algorithm can be derived. This algorithm converges to a local maximum of the approximated posterior surface. Crucially, instead of utilizing all observations, each iteration relies on a noisy yet unbiased estimate of the gradient calculated from a single randomly selected data point. Numerical simulations demonstrate the superior per-iteration computational efficiency of stochastic variational inference (SVI). While SVI typically requires more iterations for convergence, its efficiency advantage extends beyond computational speed. Albeit preliminary and somewhat speculative, the obtained results suggest that SVI yields higher-quality solutions, as evidenced by both text clustering accuracy and the implicit regularization of weakly identified components.

Keywords Dirichlet-Multinomial mixture model · Text categorization · Variational inference · Stochastic variational inference · Numerical optimization

1 Introduction

Finite mixtures of Multinomial distributions are a valuable tool for analyzing discrete positive data. This class of models has garnered considerable attention in the scientific literature, with early influential work by Nigam et al. (2000), which introduced a hierarchical structure using the Dirichlet distribution as a prior for Multinomial probabilities and demonstrated how parameter estimates from this model could be applied to the text classification of an unlabeled document corpus. In this framework, topics are represented by prior distributions over the vocabulary of terms, while the product likelihood explains why this mixture model is commonly referred to in the machine learning community as a "mixture of Unigrams." This model structure addresses limitations of pure Multinomial distributions, which tend to provide rough estimates when observed counts are close to zero and fail to adequately capture the phenomenon of word burstiness. Applying a Dirichlet prior to Multinomial probabilities is a widely accepted solution to this issue (Bouguila 2011a).

The combination of Dirichlet priors and Multinomial likelihoods yields the widely-used Dirichlet-Multinomial mixture model, which has been extended in various ways. For instance, latent topic models–of which Latent Dirichlet Allocation (LDA) is a notable example–are highly flexible and richly parameterized generative models that allow multiple topics to co-occur within a single document, as different words can be assigned to different topics (Blei et al. 2003; Blei and Lafferty 2007; Blei 2012). Despite these advancements, the simpler mixture of Unigrams remains relevant due to its inherent simplicity and superior performance in certain contexts. Specifically, for very short texts (such as abstracts, tweets, and social media posts), Blei's model often underperforms. As a consequence, Dirichlet-Multinomial mixtures have maintained their relevance in text mining and data analysis, with recent contributions focusing on efficient exploration of the posterior distribution of parameters. For example, Anderlucci and Viroli (2020) proposes an iterative method for parameter estimation in an

empirical Bayes framework, leveraging the fact that the class-conditional distributions are Dirichlet-Multinomial when the Dirichlet priors are integrated out. The authors demonstrate that their approach performs particularly well on short texts, outperforming several competing algorithms.

This semantic interpretation of our model provides a foundation for introducing the primary inferential challenge we aim to address. In scenarios involving large corpora, specifically when the number of documents n is very large (e.g. $>10^5$), the vocabulary size p representing the number of terms often becomes substantial. This remains true even after pre-processing the vocabulary to remove overly frequent terms that contribute minimal semantic value relevant to specific topics that a document may encapsulate (Manning et al. 2008). The combined dimensionality of these parameters introduces significant computational complexity, rendering standard algorithms generally non-scalable. Consequently, computation times increase drastically beyond a certain threshold of complexity, often to an unacceptable degree, despite the brevity of individual documents.

To address this challenge, Bilancia et al. (2023) propose an optimization-based algorithm for posterior parameter estimation, employing a technique known as variational inference (Jordan et al. 1999; Blei et al. 2017; Tran et al. 2021). In this approach, the joint posterior distribution is approximated by a probability distribution in which model parameters are assumed to be independent a posteriori. This setup enables the definition of a variational coordinate ascent algorithm (CAVI) for posterior parameter estimation. Optimization-based variational methods offer several advantages over traditional iterative MCMC methods. For instance, the label-switching problem–an inherent issue in finite mixture models arising from the invariance of the posterior distribution under class label permutations–makes it challenging to estimate any feature that relies on specific labels (Diebolt and Robert 1994; Celeux et al. 2000; Mena and Walker 2015). MCMC methods are sensitive to this problem, often jumping across modes in the posterior distribution, leading to outputs unsuitable for inference. In contrast, optimization-based methods focus on a single mode, which contains all the information required for posterior inference. Additionally, CAVI algorithms are generally more computationally efficient than MCMC, avoiding the extended burn-in phase and typically converging faster. However, even with CAVI, each iteration still requires using the full dataset, which poses a serious limitation when handling large-scale document corpora.

An alternative approach is to use stochastic variational inference (SVI), a scalable adaptation of the CAVI algorithm based on stochastic gradient ascent, originally introduced by Hoffman et al. (2013) (see also Bottou et al., 2018, for a comprehensive introduction to stochastic gradient-based optimization). Under certain conditions–specifically when variational distributions belong to the exponential family–the CAVI algorithm is equivalent to performing gradient ascent on a statistical manifold parameterized by the variational distribution. However, as with any coordinate gradient ascent algorithm, each data point must be reused in each iteration to re-estimate hidden variables and model parameters. In contrast, SVI relies on sampling, using a single randomly selected data point rather than the entire dataset in each iteration, resulting in a noisy but unbiased gradient estimate. Under suitable conditions, SVI converges to a local maximum in a manner similar to standard gradient ascent. Our work pro-

vides a more generalized framework than that of Hoffman et al. (2013), deriving the update equations for the method and applying them to specific cases. The only required assumptions concern the form of the complete conditionals and variational distributions, which must belong to the exponential family, with the form of each variational prior constrained by its corresponding full conditional. This foundation enables the derivation of update equations for the stochastic variational estimation of hidden variables and model parameters, demonstrating significant advantages beyond computational time alone.

The paper is organized as follows. In Sect. 2, we define the basic notation that links the model to our focus on textual data analysis. Section 3 introduces the Dirichlet-Multinomial mixture model as a hierarchical Bayesian model with latent variables, featuring an intrinsic structure in which the local context (hidden allocation variables) is juxtaposed with the global components (model parameters). In Sect. 4, we briefly discuss fixed-form variational inference and derive update equations for a coordinate ascent variational algorithm, given certain assumptions regarding the distribution of the model's full conditionals and the mean-field variational distributions. Section 5 addresses the scalability challenge and the solution provided by stochastic variational inference. The main results of the paper are presented in Sect. 6, where we derive the update equations for the stochastic variational inference of our Bayesian mixture model. Section 7 provides numerical experiments, evaluating aspects of the proposed estimation algorithm such as unsupervised text categorization, implicit regularization, and model selection. Finally, in Sect. 8, we offer conclusions and discuss future research directions.

2 Notation

As discussed in the introduction, the primary focus of the model's semantics in this paper is its application to text analysis, for which we introduce the essential notation. Let $\mathbb{V}$ represent a vocabulary of terms with $p = |\mathbb{V}|$ terms extracted from a corpus of n documents. A common assumption in this setting is that the data-generating process can be modeled as a generative probabilistic framework producing infinitely exchangeable sequences of terms. This implies that any two finite sequences of terms of the same length, differing only in the order of occurrences, are generated with the same probability and are considered identical under the bag-of-words (BOW) representation (Gelman et al. 2013). In other words, the BOW model functions as a feature generation tool, where the i-th document is represented as a vector of term counts:

$$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^\top, \tag{1}$$

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

Infinite exchangeability implies that the probability of a word occurring in $\mathbb{V}$ is independent of its position within the document. Additionally, it implies that the probability of occurrence for a finite sequence of words of arbitrary length can be factorized as the product of the corresponding marginal probabilities. These conditions charac-

terize the Unigram language model (Nigam et al. 2000), under which the likelihood of the count vector $\boldsymbol{y}_i$ for the i-th document follows the well-known Multinomial distribution:

$$p(\boldsymbol{y}_i|\boldsymbol{\beta}) = \frac{\left(\sum_{\ell=1}^{p} y_{i\ell}\right)!}{\prod_{\ell=1}^{p} y_{i\ell}!} \prod_{\ell=1}^{p} \beta_\ell^{y_{i\ell}}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the vector of Multinomial parameters that must satisfy the constraints $\beta_\ell > 0$ for $\ell = 1, \ldots, p$ and $\sum_{\ell=1}^{p} \beta_\ell = 1$.

3 Bayesian Dirichlet-Multinomial mixture models

In contrast to the standard Unigram model, we now assume the existence of k distinct probability distributions over the vocabulary, with each document in the corpus generated from exactly one of these distributions. From a Bayesian perspective, this assumption naturally gives rise to a finite Dirichlet-Multinomial mixture model, which can be formulated as the following hierarchical model:

$$\boldsymbol{y}_i|\mathbb{B}, z_{ij} = 1 \overset{\text{ind.}}{\sim} \text{Multinomial}_p(\boldsymbol{\beta}_j; y_{i+}), \quad i = 1, \ldots, n, \tag{3}$$

$$\boldsymbol{z}_i|\boldsymbol{\lambda} \overset{\text{ind.}}{\sim} \text{Multinomial}_k(\boldsymbol{\lambda}; 1), \quad i = 1, \ldots, n, \tag{4}$$

$$\boldsymbol{\beta}_j|\theta \overset{\text{ind.}}{\sim} \text{Dirichlet}_p(\mathbb{1}_p\theta), \quad j = 1, \ldots, k, \tag{5}$$

$$\boldsymbol{\lambda}|\alpha \sim \text{Dirichlet}_k(\mathbb{1}_k\alpha), \tag{6}$$

where $y_{i+} = \sum_{\ell=1}^{p} y_{i\ell}$, $\alpha, \theta > 0$, $\mathbb{1}^\top = (1, 1, \ldots, 1)$ denotes a 1-vector of suitable dimension, and $\boldsymbol{z}_i \in \mathbb{R}^k$ is a latent indicator vector defined over the canonical basis in $\mathbb{R}^k$. The index of the vector $\boldsymbol{\beta}_j^\top$ corresponds to the index of the single non-zero element in $\boldsymbol{z}_i \in \mathbb{R}^k$, such that $z_{ij} = 1$ (with all other elements being zero). For each $\boldsymbol{y}_i$, the Multinomial distribution (3) depends on y_{i+} trials. However, since this quantity appears only in the normalization constant, it can be regarded as irrelevant for posterior inference, and we will henceforth suppress it as an unnecessary nuisance parameter. Furthermore, since the vector $\boldsymbol{z}_i$ takes values in the canonical basis of $\mathbb{R}^k$, its prior distribution follows a Multinomial distribution over $z_{i+} = 1$ trials. For brevity, we will refer to this distribution as 'Multinoulli', adopting the convention suggested in Murphy (2013) by analogy with the Binomial/Bernoulli distinction, and we will denote it as $\text{Multinoulli}_k(\boldsymbol{\lambda})$, omitting the explicit indication of the number of trials in this case as well. Finally, the matrix:

$$\mathbb{B} = \begin{pmatrix} \boldsymbol{\beta}_1^\top \\ \vdots, \\ \boldsymbol{\beta}_k^\top \end{pmatrix} \in \mathbb{R}^{k \times p},$$

with the generic element being denoted as $\beta_{j\ell}$, for $j = 1, \ldots, k$ and $\ell = 1, \ldots, p$, contains k different distributions over the vocabulary of terms $\mathbb{V}$ along its rows. Each of these distributions represents a specific thematic content; however, for each document

in the corpus, it is unknown which of these distributions governs the actual data-generating process. In other words, from a semantic perspective, each document can be associated with one of k thematic contents, although the label indicating the topic is not observed. From a purely probabilistic viewpoint, the row index j in the Multinomial distribution in (3), which selects the corresponding distribution from the matrix $\mathbb{B}$, is a latent variable. This variable is modeled via the latent indicator vector $\boldsymbol{z}_i \in \mathbb{R}^k$, where $z_{ij} = 1$ and $z_{is} = 0$ for $s \neq j$. Here, $\boldsymbol{\lambda} \in \mathbb{R}^k$ represents the mixture weights. When it is necessary to make the mixture component index explicit, we can write the joint distribution of the data and the latent variables as follows:

$$p(\boldsymbol{y}_i, \boldsymbol{z}_i | \mathbb{B}, \boldsymbol{\lambda}) = \prod_{j=1}^{k} \left(\lambda_j \prod_{\ell=1}^{p} \beta_{j\ell}^{y_{i\ell}} \right)^{z_{ij}}. \tag{7}$$

The choice of a symmetric Dirichlet prior in (5) corresponds to an exchangeable prior over the Multinomial parameters $\boldsymbol{\beta}_j$, implicitly placing greater weight on the data when updating the posterior distribution of each $\boldsymbol{\beta}_j$ under a weakly informative setting of the concentration hyperparameter θ. For instance, with $\theta = 1$, the prior becomes uniform over the p-dimensional simplex.

The unnormalized posterior distribution of the latent parameters can be factorized as follows:

$$\begin{aligned} p(\mathbb{B}, \boldsymbol{z}_{1:n}, \boldsymbol{\lambda} | \boldsymbol{y}_{1:n}, \theta, \alpha) &\propto p(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \mathbb{B}, \boldsymbol{\lambda} | \theta, \alpha) \\ &= p(\boldsymbol{y}_{1:n} | \boldsymbol{z}_{1:n}, \mathbb{B}, \boldsymbol{\lambda}, \theta, \alpha) p(\mathbb{B}, \boldsymbol{z}_{1:n}, \boldsymbol{\lambda} | \theta, \alpha) \\ &= p(\boldsymbol{y}_{1:n} | \mathbb{B}, \boldsymbol{z}_{1:n}) p(\boldsymbol{z}_{1:n} | \boldsymbol{\lambda}) p(\mathbb{B} | \theta) p(\boldsymbol{\lambda} | \alpha), \end{aligned} \tag{8}$$

where $\boldsymbol{y}_{1:n} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ and $\boldsymbol{z}_{1:n}$ defined accordingly. The factorization in the last line of (8) formalizes the multivariate dependency structure among all observable and latent variables of the model, as implicitly defined in (3)–(6). However, the posterior distribution lacks a closed-form expression due to the intractability of the marginal likelihood:

$$p(\boldsymbol{y}_{1:n} | \theta, \alpha) = \prod_{i=1}^{n} \int \sum_{\boldsymbol{z}_i} p(\boldsymbol{y}_i | \mathbb{B}, \boldsymbol{z}_i) p(\boldsymbol{z}_i | \boldsymbol{\lambda}) p(\mathbb{B} | \theta) p(\boldsymbol{\lambda} | \alpha) d\mathbb{B} d\boldsymbol{\lambda}. \tag{9}$$

We must therefore rely on appropriate numerical methods for Bayesian estimation of the model parameters, keeping in mind that the chosen algorithm should scale efficiently as the corpus size n increases.

3.1 Local versus global parameters

The joint distribution of observable data and latent variables can be written as follows:

$$p(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \mathbb{B}, \boldsymbol{\lambda} | \theta, \alpha) = p(\mathbb{B} | \theta) p(\boldsymbol{\lambda} | \alpha) \prod_{i=1}^{n} p(\boldsymbol{y}_i, \boldsymbol{z}_i | \mathbb{B}, \boldsymbol{\lambda}), \tag{10}$$

where the above identity has the form:

$$p(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\delta}) = p(\boldsymbol{\omega}|\boldsymbol{\delta}) \prod_{i=1}^{n} p(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\omega}), \tag{11}$$

with $\boldsymbol{\omega} = (\mathbb{B}, \boldsymbol{\lambda})$, $\boldsymbol{\delta} = (\theta, \alpha)$ and $\mathbb{B}, \boldsymbol{\lambda}$ conditionally independent given $\boldsymbol{\delta}$. The conditional dependencies made explicit in (11) highlight the distinction between local latent variables (specific to each observed data point) and global latent variables (shared parameters). Specifically, the observed data and local latent variables are conditionally independent given the global parameters (Hoffman et al. 2013; Nguyen 2023). In our Bayesian Dirichlet-Multinomial mixture model, the local variables $\boldsymbol{z}_i$ represent the hidden cluster labels for each observation $\boldsymbol{y}_i$, while the global parameters include the mixture proportions $\boldsymbol{\lambda}$ and the Multinomial probability distributions $\mathbb{B}$ of the mixture components.

4 Fixed-form mean-field variational inference

In Bayesian finite mixture models, the well-known inferential challenges become evident when integrating out the local latent variables:

$$p(\boldsymbol{y}_i|\mathbb{B}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{z}_i} p(\boldsymbol{y}_i, \boldsymbol{z}_i|\mathbb{B}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{z}_i} p(\boldsymbol{y}_i|\mathbb{B}, \boldsymbol{z}_i) p(\boldsymbol{z}_i|\boldsymbol{\lambda}) = \sum_{j=1}^{k} \lambda_j p(\boldsymbol{y}_i|\boldsymbol{\beta}_j), \tag{12}$$

which is invariant for each of the $k!$ possible permutations of the summands.

If an exchangeable prior is assumed for the parameters, this symmetry is inherited by the posterior distribution, resulting in $k!$ symmetric modal regions that correspond to all possible permutations of the parameter labels. This phenomenon, known as 'label switching' (Celeux et al. 2000), creates substantial challenges when exploring the posterior landscape using standard MCMC sampling. Posterior sampling algorithms tend to jump between modes that differ only in the order of the labels, making it impossible to compute ergodic averages for Monte Carlo estimates of global parameters. Various solutions to this problem have been proposed, often involving constraints on the parameter space to break the symmetry of the posterior distribution, or post-processing algorithms tailored specifically for this issue. The first approach does not always eliminate all symmetries in the posterior, while the second generally performs better but introduces constraints that are not part of the prior specification, often leaving the impact on posterior estimates unclear.

In contrast, variational inference is an optimization-based approach that approximates the intractable posterior distribution with a variational distribution. For maximum generality, we will refer to the formulation in (11) and only return to the specific form of our model at a later stage. The simplest variational family of distributions is the mean-field family:

$$q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu}) = q(\boldsymbol{\omega}|\boldsymbol{\zeta}) \prod_{i=1}^{n} q(\boldsymbol{z}_i|\boldsymbol{\gamma}_i), \tag{13}$$

with $\boldsymbol{\nu} = (\boldsymbol{\zeta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$. In (13), the global parameters and local variables are assumed to be independent, with the variational parameter $\boldsymbol{\zeta}$ controlling the approximation of the posterior distribution for the global parameters, and $\boldsymbol{\gamma}_i$ governing that of the local hidden variable $\boldsymbol{z}_i$. This separation between global parameters and local variables makes approximate posterior inference computationally feasible. The quality of the approximation is then optimized by minimizing the reverse Kullback–Leibler (KL) divergence between the variational distribution and the true posterior distribution:

$$\underset{q}{\operatorname{argmin}}\, \mathsf{KL}\left(q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu}) || p(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{y}_{1:n}, \boldsymbol{\delta})\right). \tag{14}$$

Defining the evidence lower-bound (ELBO) as follows:

$$\mathsf{ELBO}(q) = \mathsf{E}_q\left[\log p(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\delta})\right] - \mathsf{E}_q\left[\log q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu})\right], \tag{15}$$

as a consequence of a standard proof using Jensen's inequality and the concavity of the logarithm function, the fundamental identity of variational inference follows (Blei et al. 2003):

$$\log p(\boldsymbol{y}_{1:n}|\boldsymbol{\delta}) = \mathsf{ELBO}(q) + \mathsf{KL}\left(q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu}) || p(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{y}_{1:n}, \boldsymbol{\delta})\right). \tag{16}$$

Since the KL divergence is always non-negative, solving the minimization problem in (14) is equivalent to identifying the member of the mean-field variational family that minimizes the reverse KL divergence from the posterior. More importantly, because the marginal log-likelihood is fixed, it follows from this relationship that minimizing the KL divergence is equivalent to maximizing the Evidence Lower Bound (ELBO). The maximized ELBO thus represents the tightest possible lower bound for the marginal log-likelihood.

4.1 Variational inference in the exponential family

Free-form variational algorithms determine the exact optimal solution for the variational distribution in (14) by setting the functional derivative to zero. In contrast, fixed-form variational methods define an explicit form for the q-distribution function, parameterized by a finite set of parameters $\boldsymbol{\nu}$ as in (13). These methods then maximize the ELBO within the $\boldsymbol{\nu}$-space, which is a subset of Euclidean space. Consequently, we adopt a broader range of assumptions to address the inferential problem effectively:

a. The full conditional of the global parameter $\boldsymbol{\omega}$ is the exponential family:

$$p(\boldsymbol{\omega}|\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta}) = h_g(\boldsymbol{\omega}) \exp\left\{\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta})^\top t(\boldsymbol{\omega}) - a_g\left(\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta})\right)\right\}, \tag{17}$$

where the natural parameter in the above equation is $\eta_g \equiv \eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta})$ and $a_g(\cdot)$ is the log-partition function. If $t(\boldsymbol{\omega})$ is a set of linearly independent functions,

the family is said to be in canonical form and $t(\boldsymbol{\omega})$ is a set of minimal sufficient statistics (Brown 1986).

b. The full conditional of each local hidden variable $\boldsymbol{z}_i$ is a canonical exponential family:

$$p(\boldsymbol{z}_i|\boldsymbol{z}_{-i}, \boldsymbol{y}_{1:n}, \boldsymbol{\omega}, \boldsymbol{\delta}) = p(\boldsymbol{z}_i|\boldsymbol{y}_i, \boldsymbol{\omega}) = h_\ell(\boldsymbol{z}_i) \exp\{\eta_\ell(\boldsymbol{y}_i, \boldsymbol{\omega})^\top t(\boldsymbol{z}_i) - a_\ell(\eta_\ell(\boldsymbol{y}_i, \boldsymbol{\omega}))\}, \quad (18)$$

with natural parameter $\eta_\ell = \eta_\ell(\boldsymbol{y}_i, \boldsymbol{\omega})$.

c. The global component of the mean field family (13) is in the same exponential family as the full conditional $p(\boldsymbol{\omega}|\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta})$:

$$q(\boldsymbol{\omega}|\boldsymbol{\zeta}) \equiv q_{\boldsymbol{\zeta}}(\boldsymbol{\omega}) = h_g(\boldsymbol{\omega}) \exp\{\boldsymbol{\zeta}^\top t(\boldsymbol{\omega}) - a_g(\boldsymbol{\zeta})\}, \quad (19)$$

while the local component is in the same exponential family as $p(\boldsymbol{z}_i|\boldsymbol{z}_{-i}, \boldsymbol{y}_{1:n}, \boldsymbol{\omega}, \boldsymbol{\delta})$:

$$q(\boldsymbol{z}_i|\boldsymbol{\gamma}_i) \equiv q_{\boldsymbol{\gamma}_i}(\boldsymbol{z}_i) = h_\ell(\boldsymbol{z}_i) \exp\{\boldsymbol{\gamma}_i^\top t(\boldsymbol{z}_i) - a_\ell(\boldsymbol{\gamma}_i)\}. \quad (20)$$

It is important to note that in (19) and (20), in order to avoid overly cumbersome notation, we have written the variational hyperparameters as if they were the natural parameters of the exponential families. However, this is not always the case, particularly for the results presented in Sect. 6. In this latter case, it is sufficient to replace $\boldsymbol{\zeta}$ with $\eta_g(\boldsymbol{\zeta})$ and $\boldsymbol{\gamma}_i$ with $\eta_\ell(\boldsymbol{\gamma}_i)$, and all the results we derive remain valid.

Hoffman et al. (2013) introduce a general framework for variational inference within exponential families, termed conditional conjugacy. This approach relies on the existence of a conditional conjugacy relationship in (11) between the prior distribution of the global variable $\boldsymbol{\omega}$ and the joint distribution $(\boldsymbol{y}_i, \boldsymbol{z}_i)$ in the local context of the ith observation, where both distributions belong to the exponential family. These conditions are essential to determine the precise form of the global natural parameter, $\eta_g \equiv \eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta})$, when applying the free-form version of variational inference. In this framework, the canonical conjugate prior for a Multinomial likelihood is utilized, specifically the Dirichlet distribution.

In our case, the assumptions in (17)–(20) are sufficient for approximating the posterior distribution using a fixed-form approach. This method is more general as it does not require the canonical conjugate prior's explicit form. Additionally, it is important to note that $\boldsymbol{\omega}$ consists of two conditionally independent components in our case. As a result, (17) must actually be represented as two distinct full conditionals, each corresponding to one parameter and independent of the other. Specifically, (17) should be expressed as the product of two conditionally independent exponential family distributions. In doing so, we recognize that when formulating the marginal posterior distribution for one of these parameters, the prior distribution of the other parameter is absorbed into the normalization constant.

We now denote the ELBO as a function of the variational parameters in the following way:

$$\mathcal{L}_{\boldsymbol{\nu}} \equiv \mathsf{ELBO}(q_{\boldsymbol{\nu}}(\boldsymbol{z}_{1:n}, \boldsymbol{\omega})), \quad (21)$$

with $q_{\boldsymbol{\nu}}(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}) \equiv q(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\nu})$, and partition the gradient vector into two sub components that correspond to the two variational vector-valued parameters:

$$\nabla_{\boldsymbol{\nu}} \mathcal{L}_{\boldsymbol{\nu}} = \left(\nabla_{\boldsymbol{\zeta}} \mathcal{L}_{\boldsymbol{\nu}}, \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\boldsymbol{\nu}}\right)^{\top}, \tag{22}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_n)$. Using assumptions (17)–(20), it is easy to show that:

$$\nabla_{\boldsymbol{\zeta}} \mathcal{L}_{\boldsymbol{\nu}} = \nabla^2_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) \left\{ \mathsf{E}_{q_{\boldsymbol{\gamma}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta}) \right] - \boldsymbol{\zeta} \right\}, \tag{23}$$

where $\nabla^2_{\boldsymbol{\zeta}}$ denotes the Hessian operator applied to the function $a_g(\boldsymbol{\zeta})$. In the same way, it can be shown that:

$$\nabla_{\boldsymbol{\gamma}_i} \mathcal{L}_{\boldsymbol{\nu}} = \nabla^2_{\boldsymbol{\gamma}_i} a_\ell(\boldsymbol{\gamma}_i) \left\{ \mathsf{E}_{q_{\boldsymbol{\zeta}}} \left[\eta_\ell(\boldsymbol{y}_i, \boldsymbol{\omega}) \right] - \boldsymbol{\gamma}_i \right\}, \quad i = 1, \ldots, n. \tag{24}$$

Proofs of these relationships are provided in Appendix A. Equations (23) and (24) suggest a straightforward coordinate-ascent variational inference algorithm (CAVI), which iteratively updates each variational parameter while keeping the others fixed at their current values. In general, it can be shown that the ELBO is concave with respect to each variational parameter individually. This implies that each maximization problem has a unique solution without requiring the computation of second partial derivatives or Hessian matrices (Plummer et al. 2020). By setting the gradients to zero, the update equations for the local variational parameters are given as follows (for $t = 0, 1, 2, \ldots$):

$$\boldsymbol{\gamma}_i^{(t+1)} = \mathsf{E}_{q_{\boldsymbol{\zeta}^{(t)}}} \left[\eta_\ell(\boldsymbol{y}_i, \boldsymbol{\omega}) \right], \quad i = 1, \ldots, n. \tag{25}$$

which are coupled with the update equation of the global variational parameter:

$$\boldsymbol{\zeta}^{(t+1)} = \mathsf{E}_{q_{\boldsymbol{\gamma}^{(t+1)}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta}) \right]. \tag{26}$$

As we will demonstrate, calculating the expected values in these equations is relatively straightforward for our hierarchical Dirichlet-Multinomial mixture model. However, updating the local variational parameters requires evaluating the function $\eta_\ell(\cdot)$ across the entire dataset. Referring to the intended semantics of our model, the global parameter $\boldsymbol{\omega}$ in (25) represents the matrix of topic probabilities, $\mathbb{B}$, where the number of columns, p, corresponds to the vocabulary size in terms $\mathbb{V}$. In large text corpora, both n (the number of observations) and p (the vocabulary size) can be substantial, and standard techniques for handling sparsity in $\mathbb{V}$ do not permit realistic scalability, as the interaction between n and p has a multiplicative effect on computational complexity (Bilancia et al. 2023). Furthermore, using a Multinomial distribution can lead to underflow issues during computation. Therefore, any method employed to operate on a logarithmic scale without underflow (e.g., the log-sum-exp trick) must also be added to the actual computational complexity. Simplifying the algorithm's structure is crucial to avoid having to process the entire dataset for each update of the local variational parameters.

5 Stochastic variational inference

Maximizing the ELBO with respect to the variational parameters $\boldsymbol{\nu}$ within a subset of Euclidean space may not be the most effective approach to ensure rapid convergence to a local maximum. The maximization of ELBO is fundamentally a functional optimization problem on a statistical manifold parameterized by the variational distribution $q_{\boldsymbol{\nu}} \equiv q_{\boldsymbol{\nu}}(\boldsymbol{z}_{1:n}, \boldsymbol{\omega})$. According to methods from information geometry, the divergence $\mathsf{KL}(q_{\boldsymbol{\nu}} \| q_{\boldsymbol{\nu}'})$ provides a suitable Riemannian metric on this manifold, conveying information about the local geometry around any fixed member of the family $q_{\boldsymbol{\nu}}$ (Amari 1998; Martens 2020). With a straightforward demonstration that avoids advanced notions of Riemannian geometry, it can be shown that the direction of maximal ascent–corresponding to the natural gradient vector on the statistical manifold–has the following approximate expression in a small neighborhood of $q_{\boldsymbol{\nu}}$ (accurate up to second order; Shrestha, 2023):

$$\nabla_{\boldsymbol{\nu}} \mathcal{L}_{\boldsymbol{\nu}}^{\mathsf{nat}} \approx \mathbb{F}_{\boldsymbol{\nu}}^{-1} \nabla_{\boldsymbol{\nu}} \mathcal{L}_{\boldsymbol{\nu}}, \tag{27}$$

where $\mathbb{F}_{\boldsymbol{\nu}}$ is the expected Fisher information matrix of the parametric family $q_{\boldsymbol{\nu}}$:

$$\mathbb{F}_{\boldsymbol{\nu}} = \mathsf{E}_{q_{\boldsymbol{\nu}}} \left[\left(\nabla_{\boldsymbol{\nu}} q_{\boldsymbol{\nu}}(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}) \right) \left(\nabla_{\boldsymbol{\nu}} q_{\boldsymbol{\nu}}(\boldsymbol{z}_{1:n}, \boldsymbol{\omega}) \right)^{\top} \right]. \tag{28}$$

For the mean-field family (13) the variational parameters are clearly orthogonal to each other:

$$\mathbb{F}_{\boldsymbol{\nu}} = \begin{pmatrix} \mathbb{F}_{\boldsymbol{\zeta}} & \mathbf{0} \\ \mathbf{0} & \mathbb{F}_{\boldsymbol{\gamma}} \end{pmatrix},$$

hence the components of the natural gradient can be analyzed separately:

$$\nabla_{\boldsymbol{\nu}} \mathcal{L}_{\boldsymbol{\nu}}^{\mathsf{nat}} \approx \begin{pmatrix} \mathbb{F}_{\boldsymbol{\zeta}} & \mathbf{0} \\ \mathbf{0} & \mathbb{F}_{\boldsymbol{\gamma}} \end{pmatrix}^{-1} \begin{pmatrix} \nabla_{\boldsymbol{\zeta}} \mathcal{L}_{\boldsymbol{\nu}} \\ \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\boldsymbol{\nu}} \end{pmatrix} = \begin{pmatrix} \mathbb{F}_{\boldsymbol{\zeta}}^{-1} \nabla_{\boldsymbol{\zeta}} \mathcal{L}_{\boldsymbol{\nu}} \\ \mathbb{F}_{\boldsymbol{\gamma}}^{-1} \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\boldsymbol{\nu}} \end{pmatrix} \approx \begin{pmatrix} \nabla_{\boldsymbol{\zeta}} \mathcal{L}_{\boldsymbol{\nu}}^{\mathsf{nat}} \\ \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\boldsymbol{\nu}}^{\mathsf{nat}} \end{pmatrix}.$$

Using the mean-field exponential family specification (19)–(20) it can be proved that (see Appendix B):

$$\mathbb{F}_{\boldsymbol{\zeta}} = \nabla_{\boldsymbol{\zeta}}^2 a_g(\boldsymbol{\zeta}), \tag{29}$$

$$\mathbb{F}_{\boldsymbol{\gamma}_i} = \nabla_{\boldsymbol{\gamma}_i}^2 a_\ell(\boldsymbol{\gamma}_i), \quad i = 1, \ldots, n, \tag{30}$$

because $\mathbb{F}_{\boldsymbol{\gamma}}$ has a block-diagonal structure, with each diagonal block being one of the matrices $\mathbb{F}_{\boldsymbol{\gamma}_i}$. Unlike the hypotheses necessary to calculate the explicit expression of the ELBO gradient, the calculation of the natural gradient obviously only requires assumptions about the form of the variational family that parameterizes the statistical manifold where we approximate the posterior distribution. Finally, by pre-multiplying the inverse of this matrix with the gradient expressions from (23) to (24), we find that:

$$\nabla_{\boldsymbol{\zeta}} \mathcal{L}_{\boldsymbol{\nu}}^{\mathsf{nat}} = \mathsf{E}_{q_{\boldsymbol{\gamma}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta}) \right] - \boldsymbol{\zeta}, \tag{31}$$

$$\nabla_{\boldsymbol{\gamma}_i} \mathcal{L}_{\boldsymbol{\nu}}^{\text{nat}} = \mathsf{E}_{q_{\boldsymbol{\xi}}} \left[\eta_\ell(\boldsymbol{y}_i, \boldsymbol{\omega}) \right] - \boldsymbol{\gamma}_i, \quad i = 1, \ldots, n. \tag{32}$$

These expressions illustrate the close relationship between the CAVI algorithm and gradient ascent for maximizing the ELBO in exponential families. Specifically, the CAVI algorithm can be interpreted as a natural gradient ascent on the statistical manifold defined by the variational distribution, assuming a unit step size. By reformulating the update equation (26) for the global variational parameter as a natural gradient ascent with a step size $\rho^{(t)}$, we arrive at (Murphy 2023):

$$\begin{aligned} \boldsymbol{\zeta}^{(t+1)} &= \boldsymbol{\zeta}^{(t)} + \rho^{(t)} \left(\mathsf{E}_{q_{\boldsymbol{\gamma}^{(t+1)}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta}) \right] - \boldsymbol{\zeta}^{(t)} \right) \\ &= (1 - \rho^{(t)}) \boldsymbol{\zeta}^{(t)} + \rho^{(t)} \mathsf{E}_{q_{\boldsymbol{\gamma}^{(t+1)}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta}) \right] \\ &= (1 - \rho^{(t)}) \boldsymbol{\zeta}^{(t)} + \rho^{(t)} \widehat{\boldsymbol{\zeta}}^{(t)}, \end{aligned} \tag{33}$$

where:

$$\widehat{\boldsymbol{\zeta}}^{(t)} = \mathsf{E}_{q_{\boldsymbol{\gamma}^{(t+1)}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta}) \right], \tag{34}$$

to emphasize that, unlike (26), the expected value of the natural parameters is not an update to the current value, but an intermediate value that must be inserted in (33).

In this formulation, no additional calculations are needed compared to the standard CAVI algorithm. However, we encounter the same issue previously discussed: updating the expected value in (34) requires updating all local variational parameters across the entire dataset. To enhance scalability, as suggested by Hoffman et al. (2013), we can approximate this expected value with a noisy but unbiased estimate that is computationally efficient. This estimate is based on a single sampled data point, $\boldsymbol{z}_s$, where:

$$s \sim \mathsf{Uniform}\,(1, 2, \ldots, n)\,,$$

and:

$$\boldsymbol{z}_s^n = (\overbrace{\boldsymbol{z}_s, \boldsymbol{z}_s, \ldots, \boldsymbol{z}_s}^{n \text{ times}}).$$

The natural gradient estimate based on this surrogate dataset is:

$$\widehat{\nabla}_{\boldsymbol{\zeta}}^{s,n} \mathcal{L}_{\boldsymbol{\nu}}^{\text{nat}} = \mathsf{E}_{q_{\boldsymbol{\gamma}}} \left[\eta_g(\boldsymbol{y}_s^n, \boldsymbol{z}_s^n, \boldsymbol{\delta}) \right] - \boldsymbol{\zeta}, \tag{35}$$

where $\boldsymbol{y}_s^n$ is defined in the same way as $\boldsymbol{z}_s^n$. Notably, when calculating the expected value, all components of $\boldsymbol{\gamma}$ with indices $\neq s$ act as multiplicative factors equivalent to the integral of the corresponding marginal component of the variational distribution, which equals 1 and thus vanishes from the expression. This allows the algorithm to operate efficiently by updating the optimized local parameter using only a single sampled data point:

$$\boldsymbol{\gamma}_s^{(t+1)} = \mathsf{E}_{q_{\boldsymbol{\zeta}^{(t)}}} \left[\eta_\ell(\boldsymbol{y}_s, \boldsymbol{\omega}) \right], \tag{36}$$

and a natural gradient ascent step based on its noisy estimate:

$$\widehat{\boldsymbol{\zeta}}^{(t)} = \mathsf{E}_{q_{\boldsymbol{\gamma}}^{(t+1)}} \left[\eta_g(\boldsymbol{y}_s^n, \boldsymbol{z}_s^n, \boldsymbol{\delta}) \right]. \tag{37}$$

which in turn must be inserted in (33).

As with other stochastic optimization algorithms where the objective function is a random unbiased estimator of the true objective, convergence to a local optimum is assured provided that the step size satisfies the Robbins–Monro conditions (Sato 2001; Kushner and Yin 2003; Blei et al. 2017; Murphy 2023):

$$\sum_{t=1}^{+\infty} \rho^{(t)} = \infty, \quad \sum_{t=1}^{+\infty} \left(\rho^{(t)}\right)^2 < \infty. \tag{38}$$

For example, one can choose (Tran et al. 2021):

$$\rho^{(t)} = (1+t)^{-\kappa}, \quad \kappa \in (0.5, 1]. \tag{39}$$

where the forgetting rate κ determines the speed at which past information is forgotten in the exponentially weighted moving average (33). As κ approaches 1, past values are downweighted more strongly, reducing their influence on the current estimate.

6 Main results

In this section, we focus on our model and its hierarchical formulation, as expressed in equation (10). We proceed by calculating the full conditional distributions and verifying whether the assumptions in Eqs. (17)–(18) hold. To this end, we define the fixed-form variational family and derive the necessary equations to explicitly formulate the update Eqs. (25) and (26).

6.1 Local variables $\boldsymbol{z}_i$

The full-conditional of local hidden indicators has expression:

$$\begin{aligned} p(\boldsymbol{z}_i|\boldsymbol{y}_{1:n}, \mathbb{B}, \boldsymbol{\lambda}, \boldsymbol{z}_{-i}, \alpha, \theta) &= p(\boldsymbol{z}_i|\boldsymbol{y}_i, \mathbb{B}, \boldsymbol{\lambda}, \alpha, \theta) \\ &\propto p(\boldsymbol{y}_i, \boldsymbol{z}_i, \mathbb{B}, \boldsymbol{\lambda}|\alpha, \theta) \\ &= p(\boldsymbol{y}_i|\boldsymbol{z}_i, \mathbb{B}) p(\boldsymbol{z}_i|\boldsymbol{\lambda}) p(\mathbb{B}|\theta) p(\boldsymbol{\lambda}|\alpha) \\ &\propto p(\boldsymbol{y}_i|\boldsymbol{z}_i, \mathbb{B}) p(\boldsymbol{z}_i|\boldsymbol{\lambda}) \\ &= p(\boldsymbol{y}_i, \boldsymbol{z}_i|\mathbb{B}, \boldsymbol{\lambda}). \end{aligned} \tag{40}$$

From (7):

$$p(\boldsymbol{y}_i, \boldsymbol{z}_i|\mathbb{B}, \boldsymbol{\lambda}) \propto \prod_{j=1}^{k} \left(\underbrace{\lambda_j \prod_{\ell=1}^{p} \beta_{j\ell}^{y_{i\ell}}}_{u_{ij}} \right)^{z_{ij}},$$

and it is obvious that in general $\sum_{j=1}^{k} u_{ij} \neq 1$, since the normalization factor of these probabilities has been absorbed into the normalization constant.

To identify the natural (non-normalized) parameters, we use the representation of this Multinomial distribution as an exponential family in non-minimal form:

$$\begin{aligned} p(\boldsymbol{y}_i, \boldsymbol{z}_i | \mathbb{B}, \boldsymbol{\lambda}) &\propto \exp\{\log p(\boldsymbol{y}_i, \boldsymbol{z}_i | \mathbb{B}, \boldsymbol{\lambda})\} \\ &= \exp\left\{ \sum_{j=1}^{k} z_{ij} \underbrace{\sum_{\ell=1}^{p} y_{i\ell} \log \beta_{j\ell}}_{h_{ij}} + \sum_{j=1}^{k} z_{ij} \log \lambda_j \right\} \\ &= \exp\left\{ \sum_{j=1}^{k} z_{ij} \left[h_{ij} + \log \lambda_j \right] \right\}, \end{aligned} \tag{41}$$

which is in the same form as (18) with:

$$\eta_\ell \equiv \eta_\ell(\boldsymbol{y}_i, \mathbb{B}, \boldsymbol{\lambda}) = \begin{pmatrix} h_{i1} + \log \lambda_1 \\ h_{i2} + \log \lambda_2 \\ \vdots \\ h_{ik} + \log \lambda_k \end{pmatrix}.$$

Since the full conditional distribution of $\boldsymbol{z}_i$ is a k-dimensional Multinomial, the variational distribution for $\boldsymbol{z}_i$ must also be Multinomial with the same dimension. Specifically, it is a Multinoulli distribution, as $\boldsymbol{z}_i$ is an indicator vector with components that sum to 1:

$$q(\boldsymbol{z}_i | \boldsymbol{\gamma}_i) \equiv q_{\boldsymbol{\gamma}_i}(\boldsymbol{z}_i) = \mathsf{Multinoulli}_k(\boldsymbol{z}_i | \boldsymbol{\gamma}_i), \quad \text{independently over } i = 1, 2, \ldots, n. \tag{42}$$

Each latent indicator $\boldsymbol{z}_i$ has its own variational parameter, allowing us to approximate the posterior distribution of each marginal component of the latent vector $\boldsymbol{z}_{1:n}$. The natural parameters of this distribution, expressed in a non-minimal exponential family form, are $\log \gamma_{ij}$ for each fixed $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Consequently, the optimal update values are given by:

$$\log \gamma_{ij} \propto \mathsf{E}_q \left[h_{ij} + \log \lambda_j \right],$$

that is:

$$\gamma_{ij} \propto \exp\left\{ \sum_{\ell=1}^{p} y_{i\ell} \mathsf{E}_q \left[\log \beta_{j\ell} \right] + \mathsf{E}_q \left[\log \lambda_j \right] \right\}. \tag{43}$$

The expression (43) cannot be further simplified, as the variational family q for the global parameters has not been fully specified. For simplicity, we have omitted the hyperparameters identifying the marginal component of q with respect to which

the expectation is taken, as this should be clear from the context. Additionally, it is evident that the local variational parameters can be normalized to 1 as follows:

$$\gamma_{ij} = \frac{\exp\left\{\mathsf{E}_q\left[h_{ij} + \log \lambda_j\right]\right\}}{\sum_{j=1}^{k} \exp\left\{\mathsf{E}_q\left[h_{ij} + \log \lambda_j\right]\right\}}. \tag{44}$$

6.2 Global parameter $\mathbb{B}$

The unnormalized full-conditional has expression:

$$\begin{aligned} p(\mathbb{B}|\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\lambda}, \alpha, \theta) &\propto p(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \mathbb{B}, \boldsymbol{\lambda}|\alpha, \theta) \\ &= p(\boldsymbol{y}_{1:n}|\boldsymbol{z}_{1:n}, \mathbb{B}) p(\boldsymbol{z}_{1:n}|\boldsymbol{\lambda}) p(\mathbb{B}|\theta) p(\boldsymbol{\lambda}|\alpha) \\ &\propto p(\boldsymbol{y}_{1:n}|\boldsymbol{z}_{1:n}, \mathbb{B}) p(\mathbb{B}|\theta) \\ &= \prod_{i=1}^{n} p(\boldsymbol{y}_i|\boldsymbol{z}_i, \mathbb{B}) p(\mathbb{B}|\theta). \end{aligned} \tag{45}$$

But it is immediate that:

$$\prod_{i=1}^{n} p(\boldsymbol{y}_i|\boldsymbol{z}_i, \mathbb{B}) \propto \prod_{i=1}^{n}\prod_{j=1}^{k}\prod_{\ell=1}^{p}\left(\beta_{j\ell}^{y_{i\ell}}\right)^{z_{ij}} = \prod_{j=1}^{k}\prod_{\ell=1}^{p}\beta_{j\ell}^{\sum_{i=1}^{n} y_{i\ell} z_{ij}},$$

whereas:

$$p(\mathbb{B}|\theta) = \prod_{j=1}^{k} p(\boldsymbol{\beta}_j|\theta) = \prod_{j=1}^{k}\prod_{\ell=1}^{p}\beta_{j\ell}^{\theta-1},$$

and therefore:

$$\prod_{i=1}^{n} p(\boldsymbol{y}_i|\boldsymbol{z}_i, \mathbb{B}) p(\mathbb{B}|\theta) \propto \prod_{j=1}^{k}\prod_{\ell=1}^{p}\beta_{j\ell}^{\sum_{i=1}^{n} y_{i\ell} z_{ij} + \theta - 1}. \tag{46}$$

This expression represents a special case of the Dirichlet-Multinomial conjugacy, as it is immediately evident that the right-hand side of the proportionality sign is the product of k unnormalized Dirichlet probability density functions, i.e.,

$$\begin{aligned} &p(\mathbb{B}|\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\lambda}, \alpha, \theta) \\ &= \prod_{j=1}^{k} \mathsf{Dirichlet}_p\left(\boldsymbol{\beta}_j \middle| \sum_{i=1}^{n} y_{i1} z_{ij} + \theta, \sum_{i=1}^{n} y_{i2} z_{ij} + \theta, \ldots, \sum_{i=1}^{n} y_{ip} z_{ij} + \theta\right). \end{aligned} \tag{47}$$

Each component of this product, written in the form of an exponential family, is proportional to:

$$\exp\left\{\log\prod_{\ell=1}^{p}\beta_{j\ell}^{\sum_{i=1}^{n}y_{i\ell}z_{ij}+\theta-1}\right\}=\exp\left\{\sum_{\ell=1}^{p}\left(\sum_{i=1}^{n}y_{i\ell}z_{ij}+\theta-1\right)\log\beta_{j\ell}\right\},$$

and comparing the above expression with (17), the components of the natural parameter are (for $j=1,\ldots,k$):

$$\eta_{gj}\equiv\eta_{gj}(\boldsymbol{y}_{1:n},\boldsymbol{z}_{1:n},\alpha,\theta)=\begin{pmatrix}\sum_{i=1}^{n}y_{i1}z_{ij}+\theta-1\\ \sum_{i=1}^{n}y_{i2}z_{ij}+\theta-1\\ \vdots\\ \sum_{i=1}^{n}y_{ip}z_{ij}+\theta-1\end{pmatrix}. \tag{48}$$

Consequently, the variational distribution of $\mathbb{B}$ must be set as the product of p-dimensional Dirichlet densities, i.e:

$$q(\boldsymbol{\beta}_j|\boldsymbol{\phi}_j)\equiv q_{\boldsymbol{\phi}_j}(\boldsymbol{\beta}_j)=\mathsf{Dirichlet}_p(\boldsymbol{\beta}_j|\boldsymbol{\phi}_j),\quad\text{independently over } j=1,\ldots,k. \tag{49}$$

The natural parameters of these distributions are $\phi_{j\ell}-1$, for $j=1,2,\ldots,k$, $\ell=1,\ldots,p$, and therefore the optimal value is:

$$\phi_{j\ell}=\mathsf{E}_q\left(\theta+\sum_{i=1}^{n}y_{i\ell}z_{ij}\right)=\theta+\sum_{i=1}^{n}y_{i\ell}\mathsf{E}_{q_{\gamma_{ij}}}(z_{ij})=\theta+\sum_{i=1}^{n}y_{i\ell}\gamma_{ij}. \tag{50}$$

6.3 Global parameter $\boldsymbol{\lambda}$

In this case:

$$\begin{aligned}p(\boldsymbol{\lambda}|\boldsymbol{y}_{1:n},\boldsymbol{z}_{1:n},\mathbb{B},\alpha,\theta)&\propto p(\boldsymbol{y}_{1:n},\boldsymbol{z}_{1:n},\mathbb{B},\boldsymbol{\lambda}|\alpha,\theta)\\&=p(\boldsymbol{y}_{1:n}|\boldsymbol{z}_{1:n},\mathbb{B})p(\boldsymbol{z}_{1:n}|\boldsymbol{\lambda})p(\mathbb{B}|\theta)p(\boldsymbol{\lambda}|\alpha)\\&\propto p(\boldsymbol{z}_{1:n}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\alpha)\\&=\prod_{i=1}^{n}p(\boldsymbol{z}_i|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\alpha),\end{aligned} \tag{51}$$

Also in this case, the explicit form of this full-conditional is a consequence of the standard Dirichlet-Multinomial conjugacy, since:

$$\prod_{i=1}^{n}p(\boldsymbol{z}_i|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\alpha)\propto\prod_{i=1}^{n}\prod_{j=1}^{k}\lambda_j^{z_{ij}}\prod_{j=1}^{k}\lambda_j^{\alpha-1}=\prod_{j=1}^{k}\lambda_j^{\alpha+\sum_{i=1}^{n}z_{ij}-1}, \tag{52}$$

from which it follows that:

$$p(\boldsymbol{\lambda}|\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \mathbb{B}, \alpha, \theta) = \mathsf{Dirichlet}_k\left(\boldsymbol{\lambda}\middle|\alpha + \sum_{i=1}^{n} z_{i1}, \alpha + \sum_{i=1}^{n} z_{i2}, \ldots, \alpha + \sum_{i=1}^{n} z_{ik}\right). \tag{53}$$

This expression, written in the form of an exponential family, becomes:

$$\begin{aligned} p(\boldsymbol{\lambda}|\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \mathbb{B}, \alpha, \theta) &\propto \exp\left\{\log \prod_{j=1}^{k} \lambda_j^{\alpha+\sum_{i=1}^{n} z_{ij}-1}\right\} \\ &= \exp\left\{\sum_{j=1}^{k}\left(\alpha + \sum_{i=1}^{n} z_{ij} - 1\right)\log \lambda_j\right\}, \end{aligned} \tag{54}$$

with:

$$\eta_g \equiv \eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \alpha, \theta) = \begin{pmatrix} \alpha + \sum_{i=1}^{n} z_{i1} - 1 \\ \alpha + \sum_{i=1}^{n} z_{i2} - 1 \\ \vdots \\ \alpha + \sum_{i=1}^{n} z_{ik} - 1 \end{pmatrix}. \tag{55}$$

Therefore, the marginal $\boldsymbol{\lambda}$-component of the variational distribution follows a Dirichlet distribution, which may not be symmetric:

$$q(\boldsymbol{\lambda}|\boldsymbol{\eta}) \equiv q_{\boldsymbol{\eta}}(\boldsymbol{\lambda}) = \mathsf{Dirichlet}_k(\boldsymbol{\lambda}|\boldsymbol{\eta}), \tag{56}$$

and since the natural parameters of this distribution are $\eta_j - 1$, it follows that the optimal values are (for $j = 1, 2, \ldots, k$):

$$\eta_j = \mathsf{E}_q\left(\alpha + \sum_{i=1}^{n} z_{ij}\right) = \alpha + \sum_{i=1}^{n} \mathsf{E}_{q_{\gamma_{ij}}}(z_{ij}) = \alpha + \sum_{i=1}^{n} \gamma_{ij}. \tag{57}$$

6.4 Stochastic variational inference for the Dirichlet-Multinomial mixture model

First, we observe that from (49), the variational distribution of $\boldsymbol{\beta}_j$ belongs to the exponential family and can be written in canonical form with minimal sufficient statistics $\log \beta_{j\ell}$, for $j = 1, \ldots, k$ and $\ell = 1, \ldots, p$. The natural parameter $\boldsymbol{\upsilon}_j$ has components $\upsilon_{j\ell} = \phi_{j\ell} - 1$. Finally, the log-partition function is given by:

$$a_g(\boldsymbol{\upsilon}_j) = \sum_{\ell=1}^{p} \log \Gamma(\upsilon_{j\ell} + 1) - \log \Gamma\left(\sum_{\ell=1}^{p}(\upsilon_{j\ell} + 1)\right)$$

$$= \sum_{\ell=1}^{p} \log \Gamma(\phi_{j\ell}) - \log \Gamma\left(\sum_{\ell=1}^{p} \phi_{j\ell}\right). \tag{58}$$

Since it is well known that the components of the gradient of the log partition function are equal to the vector of expected values of the minimal sufficient statistics (Jørgensen and Labouriau 1992):

$$\begin{aligned} \mathsf{E}_{q_{\boldsymbol{\beta}_j}}\left[\log \beta_{j\ell}\right] &= \frac{\partial a_g(\boldsymbol{\upsilon}_j)}{\partial \upsilon_{j\ell}} = \frac{\partial a_g(\boldsymbol{\upsilon}_j)}{\partial \phi_{j\ell}} \frac{\partial \phi_{j\ell}}{\partial \upsilon_{j\ell}} = \frac{\partial}{\partial \phi_{j\ell}} \log \Gamma(\phi_{j\ell}) - \frac{\partial}{\partial \phi_{j\ell}} \log \Gamma\left(\sum_{h=1}^{p} \phi_{jh}\right) \\ &= \Psi(\phi_{j\ell}) - \Psi\left(\sum_{h=1}^{p} \phi_{jh}\right), \end{aligned} \tag{59}$$

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

For the global parameter $\boldsymbol{\lambda}$, it is intuitive that similar calculations based on the variational distribution (56) lead to ($j = 1, 2, \ldots, k$):

$$\mathsf{E}_{q_{\boldsymbol{\lambda}}}\left[\log \lambda_j\right] = \Psi(\eta_j) - \Psi\left(\sum_{h=1}^{k} \eta_h\right). \tag{60}$$

The explicit expressions for these two expected values provide all the necessary information to formulate the algorithm. In contrast to the CAVI algorithm, the intermediate expected value in (37) must be computed over n replicates of the same observation and corresponding local hidden variable. For instance, the expected value in (50) is modified in the following straightforward manner:

$$\widehat{\phi}_{j\ell} = \theta + n y_{s\ell} \gamma_{sj}, \tag{61}$$

where s denotes the sampled data point. The entire processing sequence is outlined in Algorithm 1.

7 Numerical experiments

In the following, we compare the SVI and CAVI algorithms, using CAVI as the baseline to assess potential improvements. The primary comparison metric is the overall computation time, as discussed in Sects. 7.1 and 7.2. Both algorithms explore multiple modes of the surrogate posterior surface, requiring nruns replications to identify the optimal posterior mode, defined as the mode corresponding to the highest ELBO value at the end of each run (further details are provided in Sect. 7.1). For each simulation, we measured both the total computation time and the average time per run, with the latter computed by averaging over five repetitions of the entire procedure. It is important to emphasize that the results presented in this section are not only relevant in terms of computation time and determining which algorithm, SVI or CAVI, is faster, but also

Algorithm 1 Stochastic variational inference for the proposed hierarchical model

Input: Data $\boldsymbol{y}_{1:n} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$; number of components k; prior hyperparameters $\theta, \alpha > 0$; variational families $q_{\boldsymbol{\gamma}_i}(\boldsymbol{z}_i)$ for $i = 1, \ldots, n$, $q_{\boldsymbol{\phi}_j}(\boldsymbol{\beta}_j)$ for $j = 1, 2, \ldots, k$, $q_{\boldsymbol{\eta}}(\boldsymbol{\lambda})$; step size $\rho^{(t)}$ for $t = 1, 2, \ldots$

Initialize: Variational parameters γ_{ij} for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$; $\phi_{j\ell}$ for $j = 1, 2, \ldots, k$, $\ell = 1, 2, \ldots, p$; η_j for $j = 1, 2, \ldots, k$ (initialize randomly in $t = 0$).

Output: Optimized variational densities.

while the ELBO has not converged, for $t = 1, 2, \ldots$ **do**:
  Sample $s \sim \mathsf{Uniform}(1, 2, \ldots, n)$
  Update local variational parameters:
  for $j = 1, 2 \ldots, k$ **do**
    $\gamma_{sj} = \exp\left\{\sum_{\ell=1}^{p} y_{s\ell} \mathsf{E}_q\left[\log \beta_{j\ell}\right]\right\} \exp\left\{\mathsf{E}_q\left[\log \lambda_j\right]\right\}$
    $\gamma_{sj} \leftarrow \frac{\gamma_{sj}}{\sum_{h=1}^{k} \gamma_{sh}}$
  end for
  Intermediate global variational parameters:
  for $j = 1, 2, \ldots, k$ **do**
    for $\ell = 1, 2, \ldots, p$ **do**
      $\widehat{\phi}_{j\ell} = \theta + n y_{s\ell} \gamma_{sj}$
    end for
  end for
  for $j = 1, 2, \ldots, k$ **do**
    $\widehat{\eta}_j = \alpha + n\gamma_{sj}$
  end for
  Update global variational parameters:
  for $j = 1, 2, \ldots, k$ **do**
    for $\ell = 1, 2, \ldots, p$ **do**
      $\phi_{j\ell} \leftarrow \left(1 - \rho^{(t)}\right) \phi_{j\ell} + \rho^{(t)} \widehat{\phi}_{j\ell}$
    end for
  end for
  for $j = 1, 2, \ldots, k$ **do**
    $\eta_j \leftarrow \left(1 - \rho^{(t)}\right) \eta_j + \rho^{(t)} \widehat{\eta}_j$
  end for
end while

in terms of the quality of the solutions obtained. Preliminary evidence suggests that SVI, due to the inherently different geometry of the space in which the maximization occurs, yields partitions that are superoptimal compared to those obtained with CAVI. We further elaborate on this crucial point in Sect. 8.

Therefore, the second aspect of the comparison is clustering accuracy, which encompasses a set of commonly used measures for evaluating clustering performance in the literature (Fränti et al. 2024). Although the definition of clustering accuracy may initially seem counterintuitive–appearing similar to classification accuracy in supervised learning–it is important to note that clustering is an unsupervised learning task. As such, there are often no ground truth class labels, necessitating the use of internal validity measures. External validity measures, however, become relevant when ground truth clustering is available, as in our case. Among these, the adjusted Rand index (ARI) is the most widely employed (Gates and Ahn 2017). For multiclass clustering, the second external accuracy measure was obtained by calculating the best match between the true labels $c_i \in \{1, 2, \ldots, k\}$ and the predicted cluster labels $\hat{c}_i$, as follows:

$$\mathsf{Acc} = \max_{p \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(c_i = p(\hat{c}_i)\right), \tag{62}$$

where $\mathcal{P}$ is the set of all permutations in $\{1, \ldots, k\}$. To solve the optimization problem in (62) in polynomial time, we employed the Hungarian solver from the `RcppHungarian` package (Kuhn 1955; Silverman 2023) to maximize the sum of the diagonal elements of the confusion matrix, considering all permutations of rows or columns.

We would also like to emphasize that both the adjusted ARI and Acc, as defined above, are invariant to any permutation of the labels. In each run, depending on the initial conditions, the algorithm selects one of the modes of the posterior distribution, without swapping across multiple modes. These modes differ solely in the order of the labels, and such transitioning behavior, which is typical of MCMC samplers, would prevent the learning of the posterior distribution for any quantity that depends on the labels. In contrast, with variational inference, the result of the clustering quality measures remains unchanged, regardless of the order in which the labels are assigned. As such, it is entirely valid to present either the average of the best values obtained over the five repetitions of the full procedure (each with nruns runs) or one of the five results chosen randomly, to avoid distorting the results by selecting 'optimal' values that are only marginally improved by random variations. In the following, we have chosen the latter approach to simulate a real use case, where the full procedure with nruns runs is executed only once, and the best mode is then selected.

In a similar vein, Sect. 7.3 examines the semantic coherence of the produced solution, while Sect. 7.4 evaluates the performance of both algorithms as they explore the model's space with varying numbers of components. In contrast, in both Sects. 7.1 and 7.2, ARI and Acc are not employed as tools for model determination, but rather to compare the quality of the obtained partitions, conditional on a fixed value of k.

For running the variational algorithms, from relation (43) it is clear that only $\phi_{j\ell}$ and η_j need to be initialized. Following the form of (50), each element of the set $\{\phi_{j\ell};\, j = 1, \ldots, k, \ell = 1, \ldots, p\}$ was initialized as:

$$\phi_{j\ell}^{(0)} = \theta + \frac{\sum_{j=1}^{k} \sum_{\ell=1}^{p} y_{j\ell}}{k \times p} + \mathcal{N}(0, 1)_{j\ell},$$

where θ is set a priori. Similarly, each element of the set $\{\eta_j;\, j = 1, \ldots, k\}$ was initialized as (α is set a priori):

$$\eta_j^{(0)} = \alpha + \frac{n}{k} + \mathcal{N}(0, 1)_j.$$

By modifying the seed of the pseudo-random generations at each run, the algorithm ensures exploration of multiple modes of the approximate posterior surface.

7.1 Binary clustering

As an initial example, we apply stochastic variational inference to perform binary text categorization on a document corpus. The dataset used is a subset from the Reuters 21578 collection (Apté et al. 1994; Lewis et al. 2004; Debole and Sebastiani 2005), which has previously been employed by Anderlucci and Viroli (2020) and Bilancia et al. (2023) to evaluate the accuracy of other standard text clustering methods, such as hierarchical clustering and Ward's method. Our experimental corpus includes $n = 70$ documents: 50 belong to the category `acq` (covering corporate takeovers) and 20 to the category `crude` (covering crude oil news), creating a substantial class imbalance.

To convert the text into a BOW representation, we preprocessed the raw data using R 4.2.3 (R Core Team 2024) and the `tm` package (Feinerer et al. 2008; Feinerer and Hornik 2023). The preprocessing steps included the following: removing extra spaces, punctuation, and numbers; converting text to lowercase; removing stop words; applying stemming to reduce words to their base forms; and recompleting words using the most frequent match as completion. We then tokenized the text into unigrams, retaining only tokens between 4 and 16 characters in length and discarding the rest. The final output was a vocabulary of terms, $\mathbb{V}$, and a document-term matrix with dimensions $(n = 70) \times (p = 1518)$. The matrix has an overall sparsity of 96%, with an average of 54.79 words per document.

The expression for the ELBO (Evidence Lower Bound) in the Dirichlet-Multinomial mixture model was derived in Bilancia et al. (2023) for use with the standard CAVI algorithm. For completeness, we reproduce this expression in Appendix C. The same expression holds in the present context, as we use the same set of variational distributions. The only difference is the estimation algorithm employed. Algorithm convergence can be assessed by tracking the ELBO, as their values should not decrease between consecutive iterations. Any deviation from this behavior signals a programming error. However, monitoring the ELBO is also crucial for evaluating the quality of the approximation across multiple modes (Plummer et al. 2020). Evaluating it at each iteration introduces significant computational complexity into the variational algorithm. As shown in expression (C1), it is evident that the first term dominates asymptotically, resulting in an overall computational complexity of $O(n * k * p)$ for the entire run, where maxiter is a constant and does not affect the asymptotic complexity. This represents a considerable yet unavoidable additional computational burden.

In both CAVI and SVI we set $k = 2$, with default values of $\alpha = 1$ and $\theta = 5/k$. This setup is neutral regarding α and weakly informative for θ. It is also important to note that, while ELBO is concave with respect to each individual argument when others are held constant, it is generally non-concave overall. As a result, both CAVI and SVI only guarantee convergence to a local optimum, which may be sensitive to initial parameter settings (Blei et al. 2017). Variational inference typically explores a single mode of the posterior, with each run converging to a different local maximum of ELBO; hence, multiple runs are required to identify the best one. From a computational perspective, this approach can be challenging, and the solution obtained may be suboptimal, as local

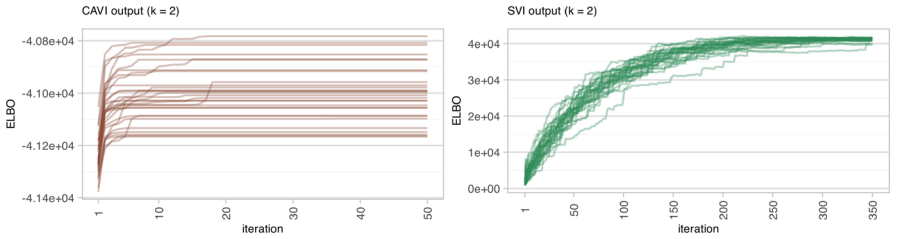

Fig. 1 Left: 30 randomly selected trajectories of the CAVI algorithm applied to a subset of the Reuters 21578 dataset with $k = 2$, `maxiter` = 50, and `nruns` = 50. To prevent the graph from becoming overly dense and visually confusing, only a subset of the trajectories was shown. Right: Same as in the left panel, but the trajectories are derived from the SVI algorithm, with `maxiter` = 350 and a forgetting rate $\kappa = 0.6$. The primary focus is on the qualitative characteristics of the graph and the comparison of convergence rates between CAVI and SVI

maxima do not necessarily include the global maximum. Figure 1 illustrates examples of ELBO trajectories toward local maxima across multiple runs.

Key considerations for interpreting the results shown in Fig. 1 include:

- Due to the inherent randomness in SVI, its convergence toward a local maximum is generally slower, and the resulting trajectories appear noisy compared to the deterministic trajectories of CAVI. All SVI-based results were obtained using a forgetting rate of $\kappa = 0.6$, though we observed substantial insensitivity in the outcomes for κ values within the range $[0.9, 0.6]$.
- SVI operates similarly to a gradient ascent with a unit step size for local parameters and adaptive step sizes, $\rho^{(t)}$, for global parameters, making the entire process effectively a minimization of $-\mathcal{L}_{\boldsymbol{v}}$. This point is noteworthy, as SVI trajectories decrease towards a local minimum, which may initially appear counterintuitive (in Fig. 1, trajectories are shown with reversed signs).
- In SVI, a single randomly selected data point provides a noisy gradient estimate. However, the actual ELBO value must be computed using the entire dataset, explaining the differences observed in Fig. 1. Both CAVI and SVI use the same observations to compute ELBO; however, CAVI has n unique values for local variational parameters, while SVI uses one value replicated n times. With random initialization, the initial additive term in ELBO (see Appendix C), which dominates, tends to be low in early iterations, so initial values of ELBO trajectories often start near zero. The two algorithms ultimately reach local maxima differing only by an additive constant.

To compare the two algorithms, we did not directly track ELBO convergence. Instead, we determined an appropriate number of iterations, denoted `maxiter`, through visual inspection to ensure convergence across all runs. We set `maxiter` = 50 for CAVI and `maxiter` = 350 for SVI. As said above, the comparison accounted for computation times and the quality of the unsupervised posterior classification, where predicted labels were obtained through the standard allocation rule for a 0/1 loss function, where $\boldsymbol{\gamma}_i^\star$ represents the optimized variational hyperparameter (James et al. 2021):

$$j(i)^{\mathsf{MAP}} = \underset{j}{\operatorname{argmax}}\, q_{\boldsymbol{\gamma}_i^\star}(z_{ij} = 1), \quad i = 1, 2, \ldots, n. \tag{63}$$

Table 1 Total, per-run, and per-iteration computation times, unsupervised classification accuracy (Acc), and the adjusted Rand index (ARI) for the two algorithms (CAVI and SVI), corresponding to multiple choices of `nruns`, where the predicted labels $\hat{c}_i$ were obtained using the allocation rule in (63)

Inference	`maxiter`	`nruns`	t (total)	t (avg. per run)	t (avg. per iter.)	Acc	ARI
CAVI	50	10	10.77 s	1.077 s	0.022 s	67.14	0.48
	50	20	21.29 s	1.065 s	0.021 s	84.29	44.36
	50	50	53.56 s	1.071 s	0.021 s	81.43	38.22
	50	100	108.47 s	1.085 s	0.021 s	92.86	72.59
SVI	350	10	33.05 s	3.305 s	0.009 s	94.29	77.26
	350	20	66.19 s	3.310 s	0.009 s	92.86	72.14
	350	50	165.23 s	3.305 s	0.009 s	97.14	88.39
	350	100	331.71 s	3.317 s	0.009 s	94.29	77.01

The results were computed on a subset of $n = 70$ documents from $k = 2$ subcategories of the Reuters 21578 collection, using an Apple Silicon M1 processor with 16 GB of RAM. For each row in the table, computation times were averaged across 5 repetitions of the overall procedure. Conversely, both Acc and ARI were derived from a randomly selected result among the five repetitions

The results are presented in Table 1. For a fixed `maxiter` value, we performed 10, 20, 50, and 100 runs (parameter `nruns`) of each algorithm to explore multiple modes of the posterior surface. For each `nruns` value, the variational parameter estimates were taken from the run where ELBO reached the highest local maximum by the end of the iterations. A preliminary observation is that the average computation time per iteration for CAVI is between 2.33 and 2.1 times higher than for SVI. Additionally, as `nruns` increases, the average iteration time for CAVI remains nearly constant, and SVI exhibits similar behavior. This is because individual runs are independent in terms of memory requirements and overall efficiency. Given that the computational complexity for each CAVI run is on the order of $O(\max(n, p) * k)$, the inefficiencies introduced by extra garbage collection and dynamic memory allocations–due to an additional nested loop traversing the entire dataset–are minimal for the current values of n and p. However, these inefficiencies would likely have a greater impact in cases where n and p are large (Bilancia et al. 2023).

Given the inherently binary structure of this experiment, we externally validate the two clustering algorithms by identifying the component associated with the `acq` corpus as the one with the highest estimated mixing weight, $\lambda_j^\star$. We then assign the label `acq` to all data points for which the value of j, determined according to the 0/1 decision rule (63), matches the previously labeled `acq` component. It is evident that this approach is entirely equivalent to computing the permuted accuracy (62). The mixing weights are estimated from the variational parameters as follows:

$$\lambda_j^\star = \frac{\eta_j^\star}{\sum_{s=1}^{k} \eta_s^\star}, \quad j = 1, 2, \ldots, k. \tag{64}$$

The obtained results demonstrate the superiority of SVI not only in computational efficiency but also in classification quality. The randomness inherent in the gradient

Table 2 The subset from Reuters 21578 used in the example, obtained by sampling with a sampling fraction of 0.1 of the $k = 5$ categories given in the first column

Category	Nr. of docs	Topic
acq	221	Corporate mergers and acquisition
crude	50	Crude oil price
earn	375	Earning reports
grain	44	Grain market and trade
money-fx	60	Foreign exchange market
Total	$n = 750$	

of SVI allows it to explore modes of the approximate posterior surface that CAVI can only reach with a substantially higher number of runs. For instance, SVI with `nruns` = 10 achieved an accuracy of 94.29% (ARI = 77.26), while with `nruns` = 50 it identified a mode with an external accuracy of 97.14% (ARI = 88.39). In contrast, achieving comparable results with CAVI required setting `nruns` = 100. In this case, computation time was approximately 328% higher than for SVI with `nruns` = 10, despite SVI requiring seven times as many iterations.

It is also noteworthy that even with `nruns` = 10, the quality of the partition produced by SVI is not inferior in terms of ARI (within decimal places) to the best partition produced by CAVI. Conversely, with `nruns` = 10, the best mode explored by CAVI has a low accuracy of 67.14 with an inconsistent ARI close to zero.

7.2 Multiclass clustering

We now evaluate the model's performance in the context of multi-class text clustering. The dataset used is an expanded subset of the Reuters 21578 collection, obtained by sampling with a 0.1 fraction from the $k = 5$ categories listed in the first column of Table 2. The pre-processing steps mirror those applied to the dataset with $k = 2$ categories, resulting in a document-term matrix of dimensions $(n = 750) \times (p = 754)$. This matrix exhibits an overall sparsity of 96%, with an average of 27.55 terms per document, consistent with the characteristics of very short texts. During preprocessing, all sparse terms $v \in \mathbb{V}$ with a document frequency $D(v) < 0.01 \times n$ were removed, where $D(v)$ denotes the number of documents in which the term v appears at least once.

Figure 2 presents results from the exploration of the modes of the surface approximating the posterior distribution. Notably, to ensure algorithm convergence, we set `maxiter` = 100 for the CAVI approach, while for SVI, we increased `maxiter` to 5000 due to the expectedly slower convergence. Further observations on these computational aspects, as well as the divergences between the two methods–attributable to differences in the underlying geometry used in the ascent to the maximum–will be discussed in Sect. 7.3.

In this case as well, the results cannot be fully interpreted without considering the efficiency with which SVI explores the modes and the quality of the solutions it

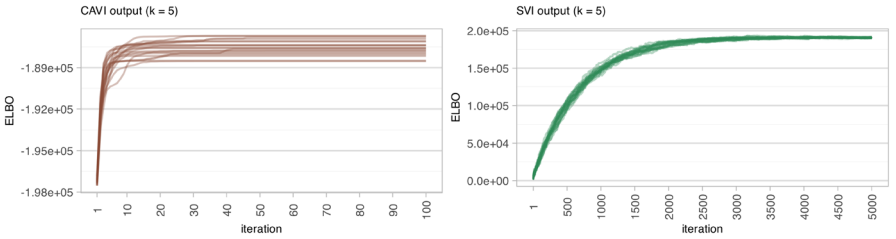

Fig. 2 "Left: 20 trajectories of the CAVI algorithm applied to a subset of Reuters 21578 data with $k = 5$, `maxiter` = 100 and `nruns` = 20. Right: Same as in the left panel, but the trajectories are derived from the SVI algorithm, with `maxiter` = 5000 and a forgetting rate $\kappa = 0.6$. The primary focus is on the qualitative characteristics of the graph and the comparison of convergence rates between CAVI and SVI

Table 3 Total, per-run, and per-iteration computation times, unsupervised classification accuracy (Acc), and the adjusted Rand index (ARI) for the two algorithms (CAVI and SVI), corresponding to multiple choices of `nruns`, where the predicted labels $\hat{c}_i$ were obtained using the allocation rule in (63). The results were computed on a subset of $n = 750$ documents from $k = 5$ subcategories of the Reuters 21578 collection, using an Apple Silicon M1 processor with 16 GB of RAM. For each row in the table, computation times were averaged across 5 repetitions of the overall procedure. Conversely, both Acc and ARI were derived from a randomly selected result among the five repetitions

Inference	`maxiter`	`nruns`	t (total)	t (avg. per run)	t (avg. per iter.)	Acc	ARI
CAVI	100	10	139.93 s	13.99 s	0.140 s	68.67	56.00
	100	20	281.19 s	14.06 s	0.141 s	60.40	40.00
	100	50	702.06 s	14.04 s	0.140 s	69.07	54.00
SVI	5000	5	1023.92 s	204.78 s	0.041 s	73.33	52.00
	5000	10	2014.56 s	201.46 s	0.040 s	73.33	52.00
	5000	15	3035.13 s	202.34 s	0.040 s	77.63	54.00
	5000	20	3941.49 s	197.07 s	0.039 s	77.65	54.00

produces. Table 3 shows that each iteration of CAVI requires approximately 3.5 times more computation time than SVI. Consequently, the best solution achieved with SVI over `nruns` = 20 modes is feasible, requiring just over one hour of CPU time.

The results obtained with CAVI guarantee an accuracy of no more than 69.07% (ARI = 54.00) with nruns = 50. In contrast, SVI consistently provides solutions that differ from those obtained with CAVI, particularly yielding higher accuracy with various choices of nruns. This suggests that these differences are unlikely to be due to chance. Specifically, with `nruns` = 20, and a total computation time approximately 5.6 times longer than that of CAVI with `nruns` = 50, SVI achieves an accuracy of 77.65% (ARI = 54.00), about 8 percentage points higher than the best accuracy achieved by CAVI.

7.3 Topic quality

In addition to performance in text classification and improvements in computation times, another important aspect is the semantic coherence of the probability distributions estimated on $\mathbb{V}$. These distributions identify low-dimensional subspaces

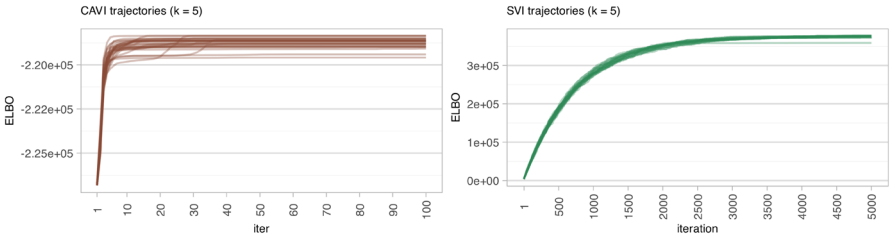

Fig. 3 Left: 30 randomly selected trajectories of the CAVI algorithm applied to the BBCsports dataset with $k = 5$, `maxiter` = 100 and `nruns` = 50. Right: Same as in the left panel, except that the trajectories were obtained using SVI with `maxiter` = 5000 and $\kappa = 0.9$ (forgetting rate)

(topics), each of which can be interpreted as carrying certain semantic content. While this construct is plausible in principle, text categorization models often produce low-dimensional subspaces that fail to be meaningful to human domain experts–i.e., low-quality topics that frequently lack coherent connections between more than a few word pairs (Meaney et al. 2023). We investigate these aspects using the BBCSport dataset, which is part of a larger collection used as a benchmark in text mining research (Greene and Cunningham 2006). This dataset contains $n = 737$ sports news articles from the BBC Sport website, published between 2004 and 2005. In this dataset, we encounter significant semantic similarity across the five available classes. Although each subset corresponds to a specific sport, all the documents share the overarching thematic content typical of news published by a sports newsroom, making coherent topic unfolding challenging. We aim to empirically test whether SVI can identify modes of the posterior surface that correspond to more interpretable topics than those generated by CAVI using the same Dirichlet-Multinomial mixture. The five subsets are:

- `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 corpus was preprocessed in the same way as in the previous examples, resulting in a large matrix of size $(n = 737) \times (p = 7883)$ with extreme sparsity of about 99%. To reduce this sparsity, we removed terms $v \in \mathbb{V}$ for which $D(v) < 0.05 \times n$. The resulting matrix was $(n = 737) \times (p = 207)$ with a sparsity of 82%. We ran CAVI with $k = 5$, `nruns` = 50, `maxiter` = 100, $\alpha = 1$, and $\theta = 5/k$, with an average time per run of 5.29 s. As shown in Fig. 3, the convergence with SVI was extremely slow. With $\alpha = 1$, $\theta = 5/k$, and $\kappa = 0.9$, we set `nruns` = 20 and `maxiter` = 5000 to ensure convergence in each run. The average time spent per run was 150.46 s, approximately 28.5 times longer than the time spent with CAVI, despite the total number of iterations per run being 50 times higher. These particularly favorable performance results in terms of computation time make the use of SVI feasible even in situations with very slow convergence, such as the present one. We will attempt to explain this phenomenon shortly.

Based on the estimates of the variational parameters, we can approximate the posterior estimates of the probability distributions $\boldsymbol{\beta}_j$ as follows:

$$\beta^{\star}_{j\ell} = \frac{\phi^{\star}_{j\ell}}{\sum_{\ell=1}^{p} \phi^{\star}_{j\ell}}, \quad j = 1, 2, \ldots, k, \quad \ell = 1, 2, \ldots, p, \tag{65}$$

and use these to identify the 10 most important words in terms of their probability of occurrence. These probabilities are plotted in Fig. 4 for each of the two algorithms. Each set is sorted based on the estimated mixing weights $\lambda^{\star}_j$ of each component (from largest to smallest). The index assigned to each component is purely conventional, as there are $k!$ equivalent modes, differing only by a permutation of the indices (though variational algorithms explore only one mode at a time):

Examining the results of CAVI, we immediately notice that the first topic (in the order of presentation) is clearly associated with terms like `chelsea` and `football`, which point directly to Football. It is also noteworthy that the estimated weight, $\lambda^{\star}_3 = 32.72\%$, closely matches the actual weight $\lambda_3 = 35.96\%$. The third component is clearly related to Athletics, as terms like `race` and `olympiad` appear, with $\lambda^{\star}_3 = 13.84\%$ compared to the actual $\lambda_3 = 13.70\%$. For the other three distributions, although the estimated weights are close to the actual weights (which are not known in real applications), we encounter significant difficulty in associating the documents classified into these groups with a well-defined thematic content. For instance, while tennis has its own specific terminology, terms with high discriminatory power (e.g., serve, let, ace, fault) are seldom mentioned in the news. These articles are often brief and primarily serve to report match results and post-match statements from participants.

In contrast, the solution based on SVI is quite different. For the first topic, the presence of the term `goal` points to Football, while `wicket` helps identify the second topic as Cricket. Similarly, the third topic is clearly related to Athletics, thanks to terms like `olympiad` and `race`. However, the fourth and fifth components reflect the weak posterior identifiability associated with a mixture of $k = 5$ components, which is not supported by the semantic structure of the data or by the weakly informative prior structure. The fourth component appears to be a repetition of the third, while the fifth contains random terms associated with a negligible mixing weight ($\lambda^{\star}_2 = 0.0037$). In other words, CAVI gives more weight to the data, providing posterior estimates similar to those obtained via maximum likelihood, but at the cost of poor topic quality and interpretability. On the other hand, SVI yields more interpretable topics (at least three) and can highlight the weak identifiability of the mixture, as two components are essentially irrelevant. Since the maximization of ELBO follows a parameterized path within the statistical manifold of variational distributions, the solution explored is likely to be flat due to weak identifiability, which also contributes to the higher number of iterations required compared to exploring directions in Euclidean parameter space. From this, we conclude that CAVI and SVI explore very different modes under certain conditions, a consequence of the different geometry of the underlying space in which they operate. Additionally, it is worth noting that $\lambda^{\star}_5 = 74.58\%$, as most of the Tennis and Rugby news are merged into this component.

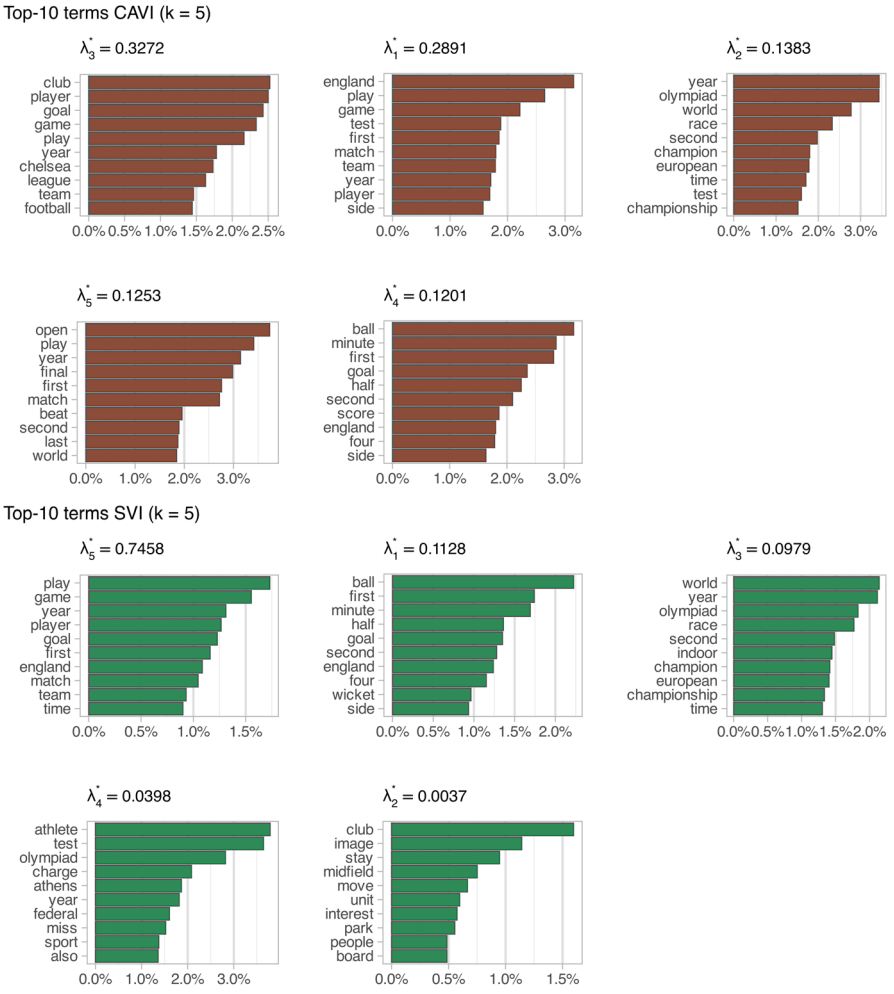

Fig. 4 Top 10 terms for CAVI and SVI of each row of the estimated $\mathbb{B}$- matrix. The rows of $\mathbb{B}^{\star}$ were sorted by the estimated weights $\lambda_j^{\star}$ $(j = 1, 2, \ldots, k)$ of each component. The index assigned to each component is purely conventional, since we have $k!$ equivalent modes that differ only up to permutation

The fact that SVI implicitly suggests the need for regularization by reducing the number of components can be formally verified using a measure of topic coherence, which can be defined for each topic $\mathbb{t}$ as: (Mimno et al. 2011; Meaney et al. 2023):

$$\mathcal{C}(\mathbb{V}^{(\mathbb{t},M)}) = \sum_{m=2}^{M} \sum_{s=1}^{m-1} \log \frac{D(v_m^{(\mathbb{t})}, v_s^{(\mathbb{t})}) + 1}{D(v_s^{(\mathbb{t})})}, \tag{66}$$

where $\mathbb{V}^{(\mathbb{t},M)} = (v_1^{(\mathbb{t})}, v_2^{(\mathbb{t})}, \ldots, v_M^{(\mathbb{t})})$ represents the list of the M most probable terms for topic $\mathbb{t}$, and $D(v, v')$ denotes the co-document frequency of the terms v and v'

Table 4 Topic coherence metric (66) for CAVI and SVI. The labels of the components are purely conventional and correspond to the order shown in Fig. 4

Inference					
CAVI					
	$\mathbb{t}=3$	$\mathbb{t}=1$	$\mathbb{t}=2$	$\mathbb{t}=5$	$\mathbb{t}=4$
$\mathcal{C}(\mathbb{V}^{(\mathbb{t},10)})$	−1.04	−0.83	−1.23	−0.93	−1.06
SVI					
	$\mathbb{t}=5$	$\mathbb{t}=1$	$\mathbb{t}=3$	$\mathbb{t}=4$	$\mathbb{t}=2$
$\mathcal{C}(\mathbb{V}^{(\mathbb{t},10)})$	−0.74	−1.16	−1.39	−1.27	−1.92

(i.e., the number of documents in which both terms v and v' occur at least once). Based on expert annotations, (Mimno et al. 2011) demonstrate the effectiveness of the measure in (66) for identifying poor topics, with lower values corresponding to topics of lower quality. The results for CAVI and SVI with $M = 10$ are shown in Table 4. For CAVI, we observe low variability around the overall average. In contrast, with SVI, we can identify the spurious component with index $j = 2$, for which $\lambda_2^\star = 0.0037$ and $C(\mathbb{V}^{(2,10)}) = -1.92$, indicating that a mixture with $k = 4$ components would be better supported in the posterior.

It is important to note that the results presented in this Subsection are quite speculative and will require further confirmation. Additionally, the approach of characterizing topics based on the estimated term distributions for each topic may be misleading. Alternative methods have been proposed for inspecting topics, which account for variations in term frequency across the corpus (Muthusami et al. 2024). Some terms might be prevalent across all topics and therefore fail to characterize a specific topic. Given these caveats, the conclusion we aim to suggest is that the SVI algorithm, operating in a completely different space from that of the CAVI algorithm, produces radically different results, even from a semantic standpoint–particularly when, as demonstrated in the example, some categories in the corpus data are not clearly separated.

7.4 Model determination

In the previous examples, the value of k was set to match the actual number of categories in the corpora, which is often unrealistic in many applications. In principle, the ELBO provides a tight approximation to the marginal log-likelihood of the model for a fixed k and could therefore be used to select the number of components when k varies. However, in practice, we do not know how tight this lower bound is, and the variational gap between ELBO and the marginal log-likelihood changes with k, complicating comparisons. Additionally, the ELBO is calculated by exploring only one mode of the approximating surface. While this is sufficient for parameter estimation, it may be invalid for model selection, where the entire approximated posterior surface must be appropriately weighted to determine the number of components (Murphy 2023). Finally, there is the issue of approximation quality (Yao et al. 2018). Looking at (C1), it is clear that the ELBO is an objective function treated as a linear sum of terms on a logarithmic scale. However, the first term, which depends on the Multinomial likelihood, has a very large negative value on this scale and dominates the other

summands. As a result, the ELBO increases slowly as k increases, leading to overfitting and sparsity in the component weights (Bilancia et al. 2023).

For these reasons, we consider an alternative, that is a slightly modified version of the BIC criterion for model selection, defined as follows (Fraley and Raftery 2002):

$$\mathsf{BIC}_k = -2\Lambda_k^\star + \mathbb{p}_k \log(n), \tag{67}$$

where $\Lambda_k^\star$ represents the log-likelihood evaluated at the posterior parameter estimates obtained from the optimal variational parameters, and $\mathbb{p}_k = k(p-1)+(k-1) = kp-1$ is the total number of free parameters. Models with lower BIC values are generally preferred. Under a set of extended regularity conditions, which include standard assumptions as well as the additional requirement that the probability distribution of the mixture components is bounded, the BIC is consistent in model selection, even for our unidentified model (Keribin 2000).

To investigate potential differences in model selection between these algorithms, we again used the BBCsport corpus due to its inherent semantic ambiguities. Specifically, we created a set of synthetic corpora by sampling from the original data as follows:

1. We set k varying in $\{3, 4, 5\}$.
2. For each given k, we randomly selected k categories from the 5 available categories (for $k = 5$, all categories were considered). For each selected category, we randomly sampled $d \in \{20, 50, 100\}$ documents, resulting in a corpus of $n = k \times d$ documents.
3. The sampling procedure, as described in 2, was repeated 50 times to obtain 50 synthetic corpora for each value of k in $\{3, 4, 5\}$.

For each corpus, we estimated the optimal variational parameters using both CAVI and SVI, and then calculated the index (67) based on these estimates. For CAVI, we set `maxiter` = 100, while for SVI, we set `maxiter` = 5000. In both cases, we used nruns = 1 to avoid exploring multiple modes, thereby focusing on model selection under possibly non-optimal conditions. Text preprocessing was performed for each corpus as described in Sect. 7.3, using the same steps and parameters. The results are presented in Table 5. From the average parameter $\overline{p}$, it is evident that the average sparsity increases significantly with higher values of k and d, ensuring that the simulation results are not overly influenced by the growing amount of information as n increases.

In line with the theoretical properties of the BIC criterion, the results of this simulation show that both CAVI and SVI tend to favor underparameterization, meaning that the number of components selected is consistently lower than the actual number of components. However, this effect appears to be more pronounced when using SVI. For example, with $k = 4$ and $n = 400$, we observe $k^\star = 3$ in 56% of the simulations and $k^\star = 4$ in 42% using CAVI, compared to $k^\star = 3$ in 62% and $k^\star = 4$ in 24% with SVI. A similar pattern emerges for $k = 5$ and $n = 500$: with CAVI, $k^\star = 4$ and $k^\star = 5$ occur in 38% and 12% of the simulations, respectively, while these percentages drop to 24% and 6% with SVI. This behavior warrants further investigation, but it is consistent with the implicit regularization effect observed in the previous example.

8 Discussion and conclusions

In this paper, we develop a stochastic variational inference framework for the Dirichlet-Multinomial finite mixture model, which has significant applications in text categorization. Our approach simplifies the original framework proposed by Hoffman et al. (2013) and demonstrates that a restricted subset of the initial assumptions is sufficient to approximate the posterior distribution of the parameters and latent allocation variables. We derive update equations for the algorithm using stochastic gradient ascent, where the gradient is replaced at each iteration by a noisy estimate based on a single, randomly selected data point. Under suitable conditions, the algorithm converges to a local maximum that depends on the initial conditions.

Due to its unique structure, this stochastic optimization algorithm does not require the entire dataset for each iteration, which significantly enhances its scalability. However, much work remains to be done. In terms of computation time, SVI outperforms CAVI per iteration by a factor of 2.8 to 3.5. Despite this, SVI generally converges more slowly than CAVI due to its inherent randomness in exploring the ascent direction. Nevertheless, it remains computationally feasible for medium-complexity datasets, owing to its scalability advantages. Moreover, SVI's efficiency is not only reflected in computation times but also in the quality of its solutions. Although SVI explores fewer modes than CAVI, there is some evidence suggesting that SVI tends to achieve systematically higher internal accuracy across identified partitions. Additionally, there is promising evidence that SVI may perform implicit regularization when the selected number k of components is not supported by the data, as some components are weakly identified with near-zero weights. The reasons behind this apparent superiority remain unclear, but they likely stem from the specific geometry of the space in which ELBO maximization occurs.

However, as shown in Tables 1 and 3, questions remain regarding the scalability of the approach for massive corpora containing millions of short texts. Substantial improvements are still needed in this area. One potential solution is to avoid monitoring the ELBO at every iteration; instead, a reasonable maximum number of iterations (`maxiter`) could be set to ensure convergence for each run, with the ELBO checked only in the final iteration to select among competing modes. Unlike the variational parameters, which are iteratively updated, the ELBO is computed at each iteration based on the current variational parameters, which–though it has certain drawbacks–significantly reduces total CPU time by several orders of magnitude. Additionally, convergence can be monitored by observing the trajectory of the variational parameters, for instance by ensuring that $\|\Phi^{(t+1)} - \Phi^{(t)}\|_{\mathbb{F}}^2 < \epsilon$ for at least h consecutive iterations, where $\Phi = \{\phi_{j\ell}\}$, $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius matrix norm, and ϵ is a predefined tolerance. Furthermore, the update of global variational parameters in Algorithm 1 relies on a series of nested for-loops, which can be optimized through vectorization. For example, Equation (61) involves multiple multiplications between the columns of the data matrix and the matrix of variational parameters, γ_{ij}. This process can be implemented more efficiently by vectorization, rather than nested for-loops (an optimization applicable to CAVI as well).

Other approaches to enhancing scalability concern the structure of the learning algorithm itself. Recent trends in scalable inference for large-scale topic models focus

Table 5 k: Actual number of categories used for sampling the BBCsport corpus. $k^\star$: Estimated number of components of the Dirichlet-Multinomial mixture, using the modified BIC criterion in (67). n: Total number of documents in each of the 50 corpora obtained by sampling the BBCSport dataset, with $n = k \times d$, $k \in \{3, 4, 5\}$ and $d \in \{20, 50, 100\}$. $\overline{p}$: Average number of terms in the 50 sampled corpora. The prior setting used is $\alpha = 1$ and $\theta = 5/k^\star$ for $k^\star$ between 2 and 7. For CAVI we used `maxiter` = 100, `nruns` = 1; for SVI, `maxiter` = 5000, `nruns` = 1 and $\kappa = 0.9$

Inference	k	d	n	$\overline{p}$	$k^\star = 2$	$k^\star = 3$	$k^\star = 4$	$k^\star = 5$	$k^\star = 6$	$k^\star = 7$
CAVI	3	20	60	273	100%	0%	0%	0%	0%	0%
	3	50	150	235	92%	8%	0%	0%	0%	0%
	3	100	300	222	18%	78%	4%	0%	0%	0%
	4	20	80	246	100%	0%	0%	0%	0%	0%
	4	50	200	224	80%	18%	2%	0%	0%	0%
	4	100	400	213	2%	56%	42%	0%	0%	0%
	5	20	100	237	100%	0%	0%	0%	0%	0%
	5	50	250	214	92%	8%	0%	0%	0%	0%
	5	100	500	205	6%	44%	38%	12%	0%	0%
SVI	3	20	60	273	100%	0%	0%	0%	0%	0%
	3	50	150	235	90%	10%	0%	0%	0%	0%
	3	100	300	222	34%	64%	2%	0%	0%	0%
	4	20	80	246	100%	0%	0%	0%	0%	0%
	4	50	200	224	92%	8%	0%	0%	0%	0%
	4	100	400	213	14%	62%	24%	0%	0%	0%
	5	20	100	237	100%	0%	0%	0%	0%	0%
	5	50	250	214	94%	6%	0%	0%	0%	0%
	5	100	500	205	20%	50%	24%	6%	0%	0%

on implementing variational inference in distributed or embarrassingly parallel environments, as discussed in Hoffman et al. (2010) and Bakhtiari and Bouguila (2014). Alternatively, online inference refers to the ability to learn incrementally from mini-batches of data, as demonstrated in Zhai et al. (2012) and Li et al. (2018). These approaches, however, differ significantly from the one explored in this work and present promising avenues for future research, potentially enabling comparisons on large-scale corpora.

Another issue pertains to the model structure itself. Under a Dirichlet prior, the mixture weight components are nearly independent, with only a slight negative correlation between pairs. However, if a corpus contains documents with an above-average frequency on a particular topic, it is realistic to expect that other related topics may also appear with above-average frequency. In contrast, the Dirichlet model enforces a repulsive interaction among components. In the context of the generative LDA model, Blei and Lafferty (2007) introduce a logistic-Normal prior over the mixing weights, which incorporates a dense, unrestricted covariance matrix. However, this prior does not fit the framework proposed by Hoffman et al. (2013) because the full conditionals do not belong to the exponential family, making inference computationally challenging. Indeed, even the full conditionals of the standard Bayesian logistic regression

model are not in the exponential family (see Blei et al., 2017; but see also Durante and Rigon, 2019 for a generalization).

Finally, another limitation of this paper lies in the absence of a comprehensive sensitivity analysis in the style of Wallach et al. (2009), as well as a systematic assessment of the influence of hyperparameter settings on the algorithm's performance. For instance, α values less than 1 tend to concentrate the posterior distribution of component weights on a subset of the vertices of the $\mathbb{R}^k$ simplex, whereas α values greater than 1 favor dense mixtures. Nevertheless, we adopted the conventional non-informative setting, which delegates most of the responsibility for updating the posterior distribution to the data. Moreover, it is important to investigate how the proposed algorithm handles overfitted mixtures–an open research question that constitutes a promising direction for future work. In this context, the results presented by Rousseau and Mengersen (2011) are particularly relevant, as they provide a systematic analysis of overfitted mixtures within the Bayesian estimation framework. Their findings demonstrate that the posterior distribution is asymptotically consistent, provided that the dimension of the mixture weight vector exceeds a critical threshold determined by the prior. Asymptotically, the superfluous components are emptied under the posterior, thereby automatically excluding irrelevant components and regularizing the posterior. It is therefore of practical importance to gain a deeper understanding of the conditions under which the prior distribution and the choice of α ensure posterior regularization and offer protection against overfitting.

A promising extension would be to use a Beta-Liouville distribution as a prior for the mixing weights $\boldsymbol{\lambda}$. This distribution generalizes the Dirichlet, allowing for a richer covariance structure in which the same expected value may correspond to different variances (Bouguila 2011b). Notably, the Beta-Liouville distribution is conjugate to the Multinomial likelihood (Bakhtiari and Bouguila 2016; Ling et al. 2022), suggesting that it may support variational inference under suitable conditions. This would help offset the increase in parameters in the Beta-Liouville model compared to the Dirichlet distribution, providing significantly more modeling flexibility. The considerations regarding the mixing weight distribution also apply to the prior distribution of the Multinomial parameters within the matrix $\mathbb{B}$. Allowing a flexible covariance structure among term frequencies could yield notable benefits when representing text as a BOW.

In conclusion, stochastic variational inference is a highly promising tool for computational inference on discrete positive data, yet there remains considerable room for improvement. Enhancements are needed both to increase computational efficiency and to extend the method's applicability to more realistic and flexible models.

Appendix A Expression of the gradient of the ELBO

For brevity, we derive only the component $\nabla_{\boldsymbol{\zeta}} \mathcal{L}_{\boldsymbol{\nu}}$ of the gradient, which corresponds to the global variational parameter, since the derivation for the local variational parameters follows a similar process. In the following, we assume the local–global structure of the model as stated in (11), and adopt the distributional assumptions outlined in (17)–(20). All constants with respect to $\boldsymbol{\zeta}$ are absorbed into an irrelevant global additive

constant. Additionally, we apply a standard result for exponential families in canonical form, which states that the components of the gradient of the log-partition function are equal to the vector of expected values of the minimal sufficient statistics (Jørgensen and Labouriau 1992):

$$\mathsf{E}_{q_{\boldsymbol{\zeta}}}\left[t(\boldsymbol{\omega})\right] = \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}),$$

where this identity is written with reference to the variational distribution of the global parameter (19). Hence:

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\nu}} &= \mathsf{E}_{q_{\boldsymbol{\nu}}}\left[\log p(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\delta}) - \log q_{\boldsymbol{\nu}}(\mathbf{z}_{1:n}, \boldsymbol{\omega})\right] \\
&= \mathsf{E}_{q_{\boldsymbol{\nu}}}\left[\log p(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\delta}) - \log q_{\boldsymbol{\zeta}}(\boldsymbol{\omega}) - \sum_{i=1}^{n} \log q_{\boldsymbol{\gamma}_i}(\mathbf{z}_i)\right] \\
&= \mathsf{E}_{q_{\boldsymbol{\nu}}}\left[\log p(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\omega}|\boldsymbol{\delta}) - \log q_{\boldsymbol{\zeta}}(\boldsymbol{\omega})\right] + \text{const.} \\
&= \mathsf{E}_{q_{\boldsymbol{\nu}}}\left[\log p(\boldsymbol{\omega}|\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta}) + \log p(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}|\boldsymbol{\delta}) - \log q_{\boldsymbol{\zeta}}(\boldsymbol{\omega})\right] + \text{const.} \\
&= \mathsf{E}_{q_{\boldsymbol{\nu}}}\left[\log p(\boldsymbol{\omega}|\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta}) - \log q_{\boldsymbol{\zeta}}(\boldsymbol{\omega})\right] + \text{const.} \\
&= \mathsf{E}_{q_{\boldsymbol{\nu}}}\Big[\log h_g(\boldsymbol{\omega}) + \eta_g(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta})^{\top} t(\boldsymbol{\omega}) - a_g(\eta_g) - \log h_g(\boldsymbol{\omega}) \\
&\quad -\boldsymbol{\zeta}^{\top} t(\boldsymbol{\omega}) + a_g(\boldsymbol{\zeta})\Big] + \text{const.} \\
&= \mathsf{E}_{q_{\boldsymbol{\nu}}}\left[\eta_g(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta})^{\top} t(\boldsymbol{\omega}) - \boldsymbol{\zeta}^{\top} t(\boldsymbol{\omega}) + a_g(\boldsymbol{\zeta})\right] + \text{const.} \\
&= \mathsf{E}_{q_{\boldsymbol{\gamma}}}\left[\eta_g(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta})\right]^{\top} \mathsf{E}_{q_{\boldsymbol{\zeta}}}\left[t(\boldsymbol{\omega})\right] - \boldsymbol{\zeta}^{\top} \mathsf{E}_{q_{\boldsymbol{\zeta}}}\left[t(\boldsymbol{\omega})\right] + a_g(\boldsymbol{\zeta}) + \text{const.} \\
&= \mathsf{E}_{q_{\boldsymbol{\gamma}}}\left[\eta_g(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta}) - \boldsymbol{\zeta}\right]^{\top} \mathsf{E}_{q_{\boldsymbol{\zeta}}}\left[t(\boldsymbol{\omega})\right] + a_g(\boldsymbol{\zeta}) + \text{const.} \\
&= \mathsf{E}_{q_{\boldsymbol{\gamma}}}\left[\eta_g(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta}) - \boldsymbol{\zeta}\right]^{\top} \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) + a_g(\boldsymbol{\zeta}) + \text{const.} \\
&= \mathsf{E}_{q_{\boldsymbol{\gamma}}}\left[\eta_g(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta})\right]^{\top} \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) - \boldsymbol{\zeta}^{\top} \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) + a_g(\boldsymbol{\zeta}) + \text{const.}
\end{aligned}$$

We now use an identity for the gradient calculation in matrix form, where $\boldsymbol{a}$ is a vector not function of $\boldsymbol{\zeta}$ (Petersen and Pedersen 2012):

$$\nabla_{\boldsymbol{\zeta}}\left(\boldsymbol{a}^{\top}\boldsymbol{u}(\boldsymbol{\zeta})\right) = \left(\nabla_{\boldsymbol{\zeta}}\boldsymbol{u}(\boldsymbol{\zeta})\right)\boldsymbol{a}.$$

Therefore:

$$\nabla_{\boldsymbol{\zeta}}\left\{\mathsf{E}_{q_{\boldsymbol{\gamma}}}\left[\eta_g(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta})\right]^{\top} \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta})\right\} = \nabla^2_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) \mathsf{E}_{q_{\boldsymbol{\gamma}}}\left[\eta_g(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}, \boldsymbol{\delta})\right].$$

In the same way, using the identity:

$$\nabla_{\boldsymbol{\zeta}}\left(\boldsymbol{u}(\boldsymbol{\zeta})^{\top}\boldsymbol{v}(\boldsymbol{\zeta})\right) = \left(\nabla_{\boldsymbol{\zeta}}\boldsymbol{u}(\boldsymbol{\zeta})\right)\boldsymbol{v}(\boldsymbol{\zeta}) + \left(\nabla_{\boldsymbol{\zeta}}\boldsymbol{v}(\boldsymbol{\zeta})\right)\boldsymbol{u}(\boldsymbol{\zeta}),$$

it follows that:

$$\nabla_{\boldsymbol{\zeta}}\left[\boldsymbol{\zeta}^{\top}\nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta})\right] = \left(\nabla_{\boldsymbol{\zeta}}\boldsymbol{\zeta}\right)\nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) + \left(\nabla^2_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta})\right)\boldsymbol{\zeta} = \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) + \left(\nabla^2_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta})\right)\boldsymbol{\zeta}.$$

Putting all this together, we obtain the expression (23):

$$\begin{aligned}\nabla_{\boldsymbol{\zeta}} \mathcal{L}_{\boldsymbol{\nu}} &= \nabla^2_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) \mathsf{E}_{q_{\boldsymbol{\gamma}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta})\right] \\ &\quad - \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) - \left(\nabla^2_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta})\right) \boldsymbol{\zeta} + \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) \\ &= \nabla^2_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) \left\{\mathsf{E}_{q_{\boldsymbol{\gamma}}} \left[\eta_g(\boldsymbol{y}_{1:n}, \boldsymbol{z}_{1:n}, \boldsymbol{\delta})\right] - \boldsymbol{\zeta}\right\}.\end{aligned}$$

It is also important to note that, in this proof, we have used the prior distribution of the global parameters in the exponential form (19), without requiring it to be the canonical conjugate prior of the Multinomial likelihood, as assumed in Hoffman et al. (2013). Finally, as already emphasized in Sect. 4, we recall that if $\boldsymbol{\zeta}$ is not the natural parameter, all derivations must be carried out with respect to the natural parameter $\boldsymbol{\eta}_{\boldsymbol{\zeta}} = \eta_g(\boldsymbol{\zeta})$, and the gradient components must be computed with respect to the components of $\boldsymbol{\eta}_{\boldsymbol{\zeta}}$. Therefore, mutatis mutandis, the proof remains valid, and the final result can subsequently be expressed in terms of the original variational parameter, as done in Sect. 6.

Appendix B The expected Fisher information matrix of the mean-field variational family

In this case, too, we give the proof for the global variational parameter $\boldsymbol{\zeta}$. The expression of the Fisher expected information matrix is a consequence of the particular structure of the variational distribution (19), which is written on a logarithmic scale as follows:

$$\log q_{\boldsymbol{\zeta}}(\boldsymbol{\omega}) = \log h_g(\boldsymbol{\omega}) + \boldsymbol{\zeta}^\top t(\boldsymbol{\omega}) - a_g(\boldsymbol{\zeta}),$$

Using the matrix identity, where $\boldsymbol{a}$ is a vector that does not depend on $\boldsymbol{\zeta}$ (Petersen and Pedersen 2012):

$$\nabla_{\boldsymbol{\zeta}} \left(\boldsymbol{\zeta}^\top \boldsymbol{a}\right) = \boldsymbol{a},$$

hence:

$$\nabla_{\boldsymbol{\zeta}} \log q_{\boldsymbol{\zeta}}(\boldsymbol{\omega}) = t(\boldsymbol{\omega}) - \nabla_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}) = t(\boldsymbol{\omega}) - \mathsf{E}_{q_{\boldsymbol{\zeta}}} \left[t(\boldsymbol{\omega})\right].$$

From this it follows that

$$\begin{aligned}\mathbb{F}_{\boldsymbol{\zeta}} &= \mathsf{E}_{q_{\boldsymbol{\zeta}}} \left[\nabla_{\boldsymbol{\zeta}} \log q_{\boldsymbol{\zeta}}(\boldsymbol{\omega}) \left(\nabla_{\boldsymbol{\zeta}} \log q_{\boldsymbol{\zeta}}(\boldsymbol{\omega})\right)^\top\right] \\ &= \mathsf{E}_{q_{\boldsymbol{\zeta}}} \left[\left(t(\boldsymbol{\omega}) - \mathsf{E}_{q_{\boldsymbol{\zeta}}} \left[t(\boldsymbol{\omega})\right]\right) \left(t(\boldsymbol{\omega}) - \mathsf{E}_{q_{\boldsymbol{\zeta}}} \left[t(\boldsymbol{\omega})\right]\right)^\top\right] \\ &= \mathsf{Cov}_{q_{\boldsymbol{\zeta}}} \left[t(\boldsymbol{\omega})\right] \\ &= \nabla^2_{\boldsymbol{\zeta}} a_g(\boldsymbol{\zeta}),\end{aligned}$$

where we have used another standard identity for exponential families in canonical form, thanks to which the Hessian matrix of the log-partition function is equal to the variance and covariance matrix of the minimal sufficient statistics (Petersen and Pedersen 2012).

Appendix C Expression of ELBO of the Dirichlet-Multinomial model

The explicit expression of ELBO for the hierarchical model (3)–(6), with mean-field variational priors given by (42), (49) and (56) was derived in Bilancia et al. (2023), which performed the standard CAVI variational inference using the gradient of ELBO in the Euclidean space. In what follows, $\boldsymbol{\nu} = (\boldsymbol{\zeta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n) \equiv (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_k, \boldsymbol{\eta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$:

$$
\begin{aligned}
\mathcal{L}_{\boldsymbol{\nu}} = & \sum_{i=1}^{n}\sum_{\ell=1}^{p}\sum_{j=1}^{k} y_{i\ell}\gamma_{ij}\left\{\Psi(\phi_{j\ell}) - \Psi\left(\sum_{\ell=1}^{p}\phi_{j\ell}\right)\right\} \\
& + \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\left\{\Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k}\eta_j\right)\right\} \\
& + k\log\Gamma(p\theta) - kp\log\Gamma(\theta) + \sum_{j=1}^{k}\sum_{\ell=1}^{p}(\theta-1)\left\{\Psi(\phi_{j\ell}) - \Psi\left(\sum_{\ell=1}^{p}\phi_{j\ell}\right)\right\} \\
& + \log\Gamma(k\alpha) - k\log\Gamma(\alpha) + \sum_{j=1}^{k}(\alpha-1)\left\{\Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k}\eta_j\right)\right\} \\
& - \sum_{j=1}^{k}\log\Gamma\left(\sum_{\ell=1}^{p}\phi_{j\ell}\right) + \sum_{j=1}^{k}\sum_{\ell=1}^{p}\log\Gamma(\phi_{j\ell}) \\
& - \sum_{j=1}^{k}\sum_{\ell=1}^{p}(\phi_{j\ell}-1)\left\{\Psi(\phi_{j\ell}) - \Psi\left(\sum_{\ell=1}^{p}\phi_{j\ell}\right)\right\} \\
& - \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\log\gamma_{ij} \\
& - \log\Gamma\left(\sum_{j=1}^{k}\eta_j\right) + \sum_{j=1}^{k}\log\Gamma(\eta_j) - \sum_{j=1}^{k}(\eta_j-1)\left\{\Psi(\eta_j) - \Psi\left(\sum_{j=1}^{k}\eta_j\right)\right\},
\end{aligned}
\tag{C1}
$$

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

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

Acknowledgements We would like to express our sincere gratitude for the support provided during the review process, as well as for the thoughtful and constructive feedback from the anonymous reviewers. We are confident that their comments and recommendations were crucial in enhancing the quality of our manuscript.

Author Contributions The authors have equally contributed.

Funding Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement. Open Access has been granted by the University of Bari Aldo Moro

Data availability The data are either publicly available or have been simulated using the R code and are reproducible.

Materials Availability Not applicable.

Code availability Available upon request.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Consent to participate Not applicable.

Consent for publication The authors have read and approved the manuscript.

Ethical approval Not applicable.

References

Amari SI (1998) Natural gradient works efficiently in learning. Neural Comput 10(2):251–276. https://doi.org/10.1162/089976698300017746

Anderlucci L, Viroli C (2020) Mixtures of Dirichlet-multinomial distributions for supervised and unsupervised classification of short text data. Adv Data Anal Classif 14(4):759–770. https://doi.org/10.1007/s11634-020-00399-3

Apté C, Damerau F, Weiss SM (1994) Automated learning of decision rules for text categorization. ACM Trans Inf Syst 12(3):233–251. https://doi.org/10.1145/183422.183423

Bakhtiari AS, Bouguila N (2014) Online learning for two novel latent topic models. In: Mahendra MS, Neuhold EJ et al (eds) Linawati. Information and communication technology. Springer, Berlin, pp 286–295

Bakhtiari AS, Bouguila N (2016) A latent Beta-Liouville allocation model. Expert Syst Appl 45:260–272. https://doi.org/10.1016/j.eswa.2015.09.044

Bilancia M, Di Nanni M, Manca F et al (2023) Variational Bayes estimation of hierarchical Dirichlet-multinomial mixtures for text clustering. Comput Stat 38(4):2015–2051. https://doi.org/10.1007/s00180-023-01350-8

Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84. https://doi.org/10.1145/2133806.2133826

Blei DM, Lafferty JD (2007) A correlated topic model of Science. Ann Appl Stat 1(1):17–35. 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(518):859–877. https://doi.org/10.1080/01621459.2017.1285773

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. SIAM Rev 60(2):223–311. https://doi.org/10.1137/16M1080173
Bouguila N (2011a) Count data modeling and classification using finite mixtures of distributions. IEEE Trans Neural Netw 22(2):186–198. https://doi.org/10.1109/TNN.2010.2091428
Bouguila N (2011b) A Liouville-based approach for discrete data categorization. In: Kuznetsov SO, Ślezak D, Hepting DH, et al (eds) Rough sets, fuzzy sets, data mining and granular computing. Lecture notes in computer science. Springer, Berlin, pp 330–337. https://doi.org/10.1007/978-3-642-21881-1_51
Brown LD (1986) Fundamentals of statistical exponential families with applications in statistical decision theory, vol 9. Institute of Mathematical Statistics. https://www.jstor.org/stable/4355554
Celeux G, Hurn M, Robert CP (2000) Computational and inferential difficulties with mixture posterior distributions. J Am Stat Assoc 95(451):957–970. https://doi.org/10.1080/01621459.2000.10474285
Debole F, Sebastiani F (2005) An analysis of the relative hardness of Reuters-21578 subsets. J Am Soc Inform Sci Technol 56(6):584–596. https://doi.org/10.1002/asi.20147
Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J Roy Stat Soc Ser B (Methodol) 56(2):363–375. https://doi.org/10.1111/j.2517-6161.1994.tb01985.x
Durante D, Rigon T (2019) Conditionally conjugate mean-field variational Bayes for logistic models. Stat Sci 34(3):472–485. https://doi.org/10.1214/19-STS712
Feinerer I, Hornik K (2023) tm: Text mining package. R package version 0.7-11. https://CRAN.R-project.org/package=tm
Feinerer I, Hornik K, Meyer D (2008) Text mining infrastructure in R. J Stat Softw 25(5):1–54. https://doi.org/10.18637/jss.v025.i05
Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631. https://doi.org/10.1198/016214502760047131
Gates AJ, Ahn YY (2017) The impact of random models on clustering similarity. J Mach Learn Res 18(87):1–28
Gelman A, Carlin J, Stern H et al (2013) Bayesian data analysis. Chapman and Hall/CRC, London
Greene D, Cunningham P (2006) Practical solutions to the problem of diagonal dominance in kernel document clustering. In: Proceedings of 23rd international conference on machine learning (ICML'06). ACM Press, pp 377–384
Hoffman M, Bach F, Blei D (2010) Online learning for latent Dirichlet allocation. In: Lafferty J, Williams C, Shawe-Taylor J, et al (eds) Advances in neural information processing systems. https://proceedings.neurips.cc/paper_files/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf
Hoffman MD, Blei DM, Wang C et al (2013) Stochastic variational inference. J Mach Learn Res 14(40):1303–1347
James G, Witten D, Hastie T et al (2021) An introduction to statistical learning, 2nd edn. Springer, New York
Jordan MI, Ghahramani Z, Jaakkola TS et al (1999) An introduction to variational methods for graphical models. Mach Learn 37(2):183–233. https://doi.org/10.1023/A:1007665907178
Jørgensen B, Labouriau RS (1992) Exponential families and theoretical inference, Monogr. Mat., Rio de Janeiro, vol 52. Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro
Keribin C (2000) Consistent estimation of the order of mixture models. Sankhyā: Indian J Stat Ser A (1961–2002) 62(1):49–66
Kuhn HW (1955) The Hungarian method for the assignment problem. Naval Res Logist Q 2(1–2):83–97. https://doi.org/10.1002/nav.3800020109
Kushner HJ, Yin GG (2003) Stochastic approximation and recursive algorithms and applications, stochastic modelling and applied probability, vol 35, 2nd edn. Springer, New York
Lewis DD, Yang Y, Rose TG et al (2004) RCV1: a new benchmark collection for text categorization research. J Mach Learn Res 5:361–397
Li Y, Song WZ, Yang B (2018) Stochastic variational inference-based parallel and online supervised topic model for large-scale text processing. J Comput Sci Technol 33(5):1007–1022. https://doi.org/10.1007/s11390-018-1871-y
Ling Y, Guan W, Ruan Q et al (2022) Variational learning for the inverted beta-Liouville mixture model and its application to text categorization. Int J Interact Multimed Artif Intell 7(5):76. https://doi.org/10.9781/ijimai.2022.08.006
Ma S (2001) Online model selection based on the variational Bayes. Neural Comput 13(7):1649–1681. https://doi.org/10.1162/089976601750265045

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, New York

Martens J (2020) New insights and perspectives on the natural gradient method. J Mach Learn Res 21(146):1–76

Meaney C, Stukel TA, Austin PC et al (2023) Quality indices for topic model selection and evaluation: a literature review and case study. BMC Med Inform Decis Mak 23(1):132. https://doi.org/10.1186/s12911-023-02216-1

Mena RH, Walker SG (2015) On the Bayesian mixture model and identifiability. J Comput Graph Stat 24(4):1155–1169. https://doi.org/10.1080/10618600.2014.950376

Mimno D, Wallach H, Talley E et al (2011) Optimizing semantic coherence in topic models. In: Barzilay R, Johnson M (eds) Proceedings of the 2011 conference on empirical methods in natural language processing. Association for computational linguistics, Edinburgh, Scotland, pp 262–272. https://aclanthology.org/D11-1024

Murphy KP (2013) Machine learning: a probabilistic perspective. The MIT Press, Cambridge

Murphy KP (2023) Probabilistic machine learning: an introduction. MIT Press, Cambridge

Muthusami R, Mani Kandan N, Saritha K et al (2024) Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. Sci Rep 14(1):12003. https://doi.org/10.1038/s41598-024-61738-4

Nguyen D (2023) An in depth introduction to variational Bayes note. https://papers.ssrn.com/abstract=4541076

Nigam K, Mccallum AK, Thrun S et al (2000) Text classification from labeled and unlabeled documents using EM. Mach Learn 39(2/3):103–134. https://doi.org/10.1023/A:1007692713085

Petersen KB, Pedersen MS (2012) The matrix cookbook. version 20121115. http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html

Plummer S, Pati D, Bhattacharya A (2020) Dynamics of coordinate ascent variational inference: a case study in 2D Ising models. Entropy 22(11):1263. https://doi.org/10.3390/e22111263

R Core Team (2024) R: a language and environment for statistical computing. version 4.2.3. https://www.R-project.org/

Rousseau J, Mengersen K (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. J R Stat Soc Ser B (Stat Methodol) 73(5):689–710. https://doi.org/10.1111/j.1467-9868.2011.00781.x

Shrestha R (2023) Natural gradient methods: perspectives, efficient-scalable approximations, and analysis. http://arxiv.org/abs/2303.05473

Sieranoja S, Fränti P, Fränti P et al (2024) Clustering accuracy. Appl Comput Intell 4(1):24–44. https://doi.org/10.3934/aci.2024003

Silverman J (2023) RcppHungarian: solves minimum cost bipartite matching problems. R package version 0.3. https://CRAN.R-project.org/package=RcppHungarian

Tran MN, Nguyen TN, Dao VH (2021) A practical tutorial on variational Bayes. http://arxiv.org/abs/2103.01327

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

Yao Y, Vehtari A, Simpson D et al (2018) Yes, but did it work?: Evaluating variational inference. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, vol 80. PMLR, pp 5581–5590. https://proceedings.mlr.press/v80/yao18a.html

Zhai K, Boyd-Graber J, Asadi N et al (2012) Mr. LDA: a flexible large scale topic modeling package using variational inference in MapReduce. In: Proceedings of the 21st international conference on World Wide Web. Association for Computing Machinery, New York, WWW '12, pp 879–888. https://doi.org/10.1145/2187836.2187955