Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-026-00690-9

REGULAR ARTICLE

Sub-topics detection with extended flexible latent Dirichlet allocation

Roberto Ascari[1] · **Alice Giampino**[1] · **Sonia Migliorati**[1]

Received: 29 November 2025 / Revised: 18 May 2026 / Accepted: 26 May 2026

Abstract

The rapid expansion of digital data has intensified the need for computational methods capable of analyzing complex latent structures across a variety of domains, including textual data. Latent topic models, particularly latent Dirichlet allocation (LDA), are widely used to uncover latent structures in large text corpora. However, the Dirichlet prior on topic proportions imposes structural limitations that reduce the model's ability to capture complex dependencies among topics. In this paper, we introduce the extended flexible latent Dirichlet allocation (EFLDA), a probabilistic model that extends LDA by allowing richer patterns of dependence among topics. The enriched parametrization of EFLDA improves the model's ability to represent complex thematic structures, leading to great interpretability in real-world settings. Furthermore, we introduce the concept of sub-topics, defined as specific combinations of topics that provide a deeper understanding of corpora. We develop a collapsed Gibbs sampler for efficient inference and conduct an extensive evaluation on both synthetic data and multiple real-world applications, including mental health discourse, news articles, and microbiome data. Empirical results show that EFLDA outperforms classical LDA and recent alternative approaches in terms of topic coherence, sub-topic detection, and interpretability, while remaining robust across heterogeneous data settings characterized by complex and overlapping latent structures.

Keywords Collapsed Gibbs sampling · Finite mixture · Latent variables · Probabilistic modeling · Topic models

✉ Roberto Ascari
roberto.ascari@unimib.it

Alice Giampino
alice.giampino@unimib.it

Sonia Migliorati
sonia.migliorati@unimib.it

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

Published online: 01 July 2026

Mathematics Subject Classification 62H30 · 62P10 · 62P10

1 Introduction

The exponential growth of digital textual data observed in recent years has provided new opportunities to study complex phenomena across a wide range of domains, including social sciences, digital health, and biological systems, through computational approaches. In this regard, latent topic models have emerged as interesting probabilistic frameworks for uncovering latent thematic structures within large collections of unstructured data. Among these, the latent Dirichlet allocation (LDA, Blei et al. 2003) emerged as the most widespread and popular. By identifying coherent clusters of words and documents, topic models enable researchers to systematically explore patterns of communication, attitudes, and behaviors across diverse settings, spanning public health, digital discourse, and journalistic archives, through to non-textual datasets such as microbiome profiles, as well as to detect trends and headline topics in specific contexts.

Despite the methodological advances in topic models, the original LDA formulation remains the most widely used reference model in the field. However, LDA has well-known limitations arising from the rigidity of the Dirichlet prior imposed on the topic proportions. Indeed, the Dirichlet distribution imposes a strong constraint: topic proportions of a specific document are *nearly* independent (Aitchison 2003), implying that the prevalence of one topic in a document is independent (in some sense proper for compositional vectors) of that of another. This assumption is often unrealistic, as topics in real-world corpora can be correlated. The limited parameterization of the Dirichlet prevents it from adequately modeling such dependencies. In particular, the off-diagonal entries of the Dirichlet covariance or correlation matrices are always negative and proportional to the product of the expected topic proportions, which constrains the model's ability to represent richer relationships among topics. Consequently, developing more flexible latent topic models that can represent more general correlations among topics and account for overdispersion remains an open challenge in the analysis of both textual documents and other high-dimensional count data.

In this paper, we propose a new latent topic model, called extended flexible latent Dirichlet allocation (EFLDA), which relaxes the constraints of the Dirichlet prior and enables richer dependencies among topics through a more flexible prior specification. The model can impose positive topic correlations, thus overcoming fundamental limitations of the traditional LDA formulation. The richer parametrization further accommodates scenarios where topics exhibit complex dependence patterns. This flexibility also allows us to define *sub-topics*, namely coherent topic variations nested within broader themes, supporting a more detailed interpretation of latent structures across heterogeneous data types. Sub-topics introduce additional structure by modeling topic-specific variability that existing LDA-based approaches cannot capture. Empirically, this increased flexibility translates into improved performance with respect to competing approaches, as demonstrated in three real data applications, spanning different domains. In particular, EFLDA consistently achieves competitive

or superior results while preserving a fully probabilistic and interpretable framework, highlighting its robustness across heterogeneous settings.

The paper is organized as follows. Section 1.1 reviews the literature related to the topic modeling approaches and their applications. Section 2 introduces the general structure of topic models, with a particular focus on LDA and the proposed EFLDA, and discusses their probabilistic formulations. Section 3 describes the collapsed Gibbs sampling (CGS) algorithm used for posterior inference. Section 4 describes a procedure to detect and label sub-topics. Section 5 outlines the experimental design employed to develop an initialization strategy, whereas Sect. 6 presents empirical results and comparisons with competing models using mental health counseling conversations, microbiome data, and newspaper sport articles from BBC. Finally, Sect. 7 offers concluding remarks and future research directions. Additional proofs, computational details, figures, and tables are provided in the Supplementary Material (SM).

1.1 Related works

Latent topic models have been widely applied across a variety of domains, ranging from social media and public health to more traditional text corpora. In this context, news articles represent a primary and well-structured source of textual data for latent modeling (Feuerriegel et al. 2016; Ahmed et al. 2022; Griciūtė et al. 2023), as they can be systematically retrieved from extensive digital archives and databases. In particular, benchmark datasets derived from news collections, including the BBC Sport dataset (Greene et al. 2014), are commonly used to evaluate the ability of topic models to recover coherent and interpretable thematic structures. These datasets provide well-defined categories, such as sports, politics, and business, and represent a useful testbed for assessing both model interpretability and clustering performance. The sports domain, in particular, offers a controlled setting with relatively clear thematic boundaries, while still exhibiting variability in language and narrative style, making it suitable for comparing different topic modeling approaches.

Early applications in public health demonstrated the utility of topic modeling for large-scale surveillance and the extraction of meaningful health-related discourse. Paul and Dredze (2014) employed LDA on Twitter data to automatically identify health topics that exhibited significant correlations with validated public health indicators, thereby illustrating the feasibility of minimally supervised approaches for real-world health data analysis. Subsequent studies extended this framework to clinical and epidemiological domains: Bittermann and Fischer (2018) and Liu et al. (2021) applied topic modeling to detect emergent research trends in the clinical psychology literature, while Xue et al. (2020) analyzed psychological reactions and public discourse on social media during the COVID-19 pandemic. A recent work has combined topic modeling with word embeddings to analyze narratives of violent deaths, enabling the identification of interpretable topics that are often not captured by predefined variables (Arseniev-Koehler et al. 2022).

In the domain of mental health, latent topic models have facilitated the exploration of the psychological and social dimensions embedded in online communication. Jones et al. (2019) demonstrated that Reddit users at risk for suicide tend to discuss distinct thematic patterns compared to non-suicidal users, suggesting that topic struc-

tures may capture linguistic markers of psychological vulnerability. Similarly, Carron-Arthur et al. (2016) and Dao et al. (2017) utilized LDA to analyze online support communities, yielding large-scale quantitative descriptions of user-generated content and elucidating the diversity of mental health concerns expressed within digital peer networks. Yin et al. (2022) further investigated mental health attitudes through analyses of posts on the Chinese equivalent of Twitter, highlighting the relationship between online discourse and broader sociocultural determinants of psychological well-being. Reddit posts have also been utilized for the early detection of psychological disorders through temporal word embeddings (Couto et al. 2025), which facilitate the study of linguistic evolution over time. Furthermore, such social media data have been employed to test whether narrowing the scope of topic modeling to predefined medical concepts and semantic types provides a more robust framework for interpreting unstructured digital discourse (Xin et al. 2025).

Beyond social media data, topic models have also been applied to open-ended survey responses, addressing the analytical challenges inherent in the interpretation of qualitative data. Finch et al. (2018) employed latent topic models to extract underlying themes from respondents' written narratives, thereby overcoming limitations associated with traditional Likert-type response formats. In a related line of research, Westrupp et al. (2022) used topic modeling to identify key issues faced by parents in supporting children's mental health, emphasizing the relevance of early detection and targeted intervention strategies. In the context of clinical conversations, Salmi et al. (2024) demonstrated that standard topic modeling often struggles with the high variability of local dialogue. Their findings suggest that analyzing data at the utterance or segment level, rather than treating entire conversations as single documents, significantly improves coherence. This is particularly effective when leveraging transformer-based embeddings and density-based clustering to isolate and filter out non-informative or noisy dialogue. Complementing these applied studies, Gao and Sazara (2023) conducted a large-scale review of topic modeling research in mental health, underscoring the method's growing prominence and its capacity to reveal high-impact research trends in the field. Another recent review by Hagg et al. (2022) examined the application of LDA within psychological research, noting its popularity alongside a concerning sensitivity to analytical parameters. The authors highlighted that reproducibility remains a significant challenge due to inconsistent reporting standards, particularly in social media contexts. To address these gaps, they introduced a "LDA preferred reporting checklist", designed to standardize the documentation of key methodological choices. Labeling and interpretation remain further critical challenges in topic modeling. Addressing this, Mekaoui et al. (2025) provided a comprehensive systematic review of 41 studies, offering a rigorous taxonomy of existing labeling techniques and establishing a clearer framework for researchers navigating this issue.

Over the past two decades, a variety of extensions and generalizations of LDA have been proposed (see Jelodar et al. (2019) for a review). For instance, models have been developed to handle zero-inflated term frequencies (Deek et al. 2021), to incorporate hierarchical or correlated topic structures (Blei and Lafferty 2007), or to capture document-level variations in topic composition (LeBlanc et al. 2023). Latent topic models have also gained popularity in settings where the primary focus is not

on textual data. For instance, several studies have applied topic models to microbiome data, in which they are often referred to as mixed-membership models (Higashi et al. 2018; Sankaran and Holmes 2019; Hosoda et al. 2020; Breuninger et al. 2021; Deek et al. 2021; LeBlanc et al. 2023; Giampino et al. 2025). In this context, the observed count matrix of bacterial taxa for each subject is treated analogously to the data representation used in textual applications.

Recent advances have explored alternative representations of the latent space to improve topic separability and document clustering. In particular, Schiavon (2025) introduced a novel topic modeling framework based on reduced latent space clustering, enabling efficient document categorization while preserving meaningful semantic structure. In a similar direction, recent embedding-based approaches, such as BERTopic (Grootendorst 2022) and Top2Vec (Angelov 2020), leverage pretrained language models to derive dense semantic representations of documents, which are then grouped using clustering algorithms to automatically identify coherent topics. BERTopic combines transformer-based embeddings with dimensionality reduction and density-based clustering, followed by class-based term weighting to produce interpretable topic representations. Differently, Top2Vec jointly learns document and word embeddings and identifies topics as dense regions in the embedding space. Both BERTopic and Top2VEC perform topic discovery without requiring the number of topics to be specified a priori. Thus, these methods represent a compelling alternative to probabilistic latent topic models when corpora exhibit a meaningful semantic structure. In such settings, vector representations effectively capture semantic relationships among documents, facilitating the recovery of coherent and interpretable topics. Conversely, in the absence of a clear semantic structure (e.g., in the microbiome framework), these methods may face substantial limitations. Indeed, in these settings, the corpus can essentially be reduced to a document-term count matrix, where each entry represents the frequency of a given word within a document. As a consequence, word ordering carries little or no semantic information, and different permutations of the corpus may lead to significantly different results, both in terms of the number of detected topics and their internal composition. Even under identical model configurations, the resulting topic structures may vary considerably, highlighting an intrinsic instability of these approaches when applied to semantically weak or poorly structured corpora.

Moreover, comparing embedding-based approaches with probabilistic topic models raises an additional methodological issue related to text preprocessing. Classical probabilistic models typically rely heavily on preprocessing steps, such as stopword removal or stemming, whereas embedding-based methods are generally designed to operate directly on raw text. Consequently, a fair comparison requires deciding whether preprocessing should be applied uniformly across methods. On the one hand, avoiding preprocessing may result in models being fitted on substantially different corpora; on the other hand, applying preprocessing may partially alter the semantic structure that embedding-based methods are intended to exploit.

Another widely used approach for extracting latent topics is non-negative matrix factorization (NMF, Kuang et al. 2014), a non-probabilistic matrix decomposition method that enforces non-negativity constraints. While these methods effectively capture semantic relationships through embedding spaces, they typically lack an

explicit probabilistic generative structure. In contrast, probabilistic models, such as LDA, FLDA, and the proposed EFLDA, rely on a well-defined generative framework in which documents are modeled as mixtures of topics and topics as distributions over words. This enables explicit probabilistic interpretation of topic–document and topic–word distributions, supports uncertainty quantification and principled statistical inference. For these reasons, probabilistic approaches remain highly relevant, particularly in applications where interpretability and statistical inference are of primary importance. Within this framework, the EFLDA introduces greater flexibility in modeling topic structure, providing a novel tool for this class of analysis.

2 Topic models

From a statistical perspective, latent topic models provide a structured representation of a corpus of D textual documents, each consisting of N_d words, through two sets of discrete distributions. The first set consists of K distributions over the *vocabulary*, namely, the set $\mathcal{V}$ containing the V unique words observed in the corpus, and is referred to as the topics. The second set captures the relative importance of each topic within individual documents. These distributions can be represented as vectors lying on appropriate simplices, specifically the set $\mathcal{S}_P = \left\{ \mathbf{x} = (x_1, \ldots, x_P)^\intercal : x_p > 0, \sum_{p=1}^{P} x_p = 1 \right\}$.

More precisely, on the one hand, the vectors $\boldsymbol{\phi}_k \in \mathcal{S}_V$ define the k-th topic, for $k = 1, \ldots, K$, where each element $\phi_{k,v}$ denotes the probability assigned by topic k to the v-th word in the vocabulary, $v \in \{1, \ldots, V\}$. Typically, distinct topics allocate different (non-zero) probabilities to each word. On the other hand, documents are represented by vectors $\boldsymbol{\theta}_d \in \mathcal{S}_K$, $d = 1, \ldots, D$, where $\theta_{d,k}$ quantifies the proportion of words in document d arising from topic k and K is the number of topics assumed by the model.

Latent topic models typically rely on a key assumption known as *bag-of-words* (Blei et al. 2003). This assumption, which can be interpreted as a form of exchangeability of words within documents, implies that the order of the words within a document does not provide any additional information, since only the frequency of each word is considered relevant. As a result, a corpus can be summarized by a $D \times V$ document-term matrix (DTM), where each entry (d, v) denotes the number of times the v-th word in the vocabulary appears in the d-th document.

By relying on the *bag-of-words*, latent topic models may be used to create corpora through the following simple corpus-generating mechanism:

i) Generate the topic vectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ independently from a distribution defined on $\mathcal{S}_V$.
ii) Generate the vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D$ representing the D documents independently from a distribution defined on $\mathcal{S}_K$.
iii) For each document $d \in \{1, \ldots, D\}$ and for each word $n \in \{1, \ldots, N_d\}$:

    1. draw a topic label $z_{d,n}$ from $Z_{d,n} \sim \text{Categorical}(\boldsymbol{\theta}_d)$;

2. draw a word $w_{d,n} \in \mathcal{V}$ from $W_{d,n}|z_{d,n} = k \sim \text{Categorical}(\boldsymbol{\phi}_k)$.

Different latent topic models are defined by imposing different distributions on the parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$. For example, the LDA (Blei et al. 2003), one of the most widespread latent topic models, is defined by assuming that both $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ are distributed according to Dirichlet distributions: $\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha})$ and $\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\beta})$, where $\boldsymbol{\alpha} \in \mathbb{R}_+^K$ and $\boldsymbol{\beta} \in \mathbb{R}_+^V$ are vectors on non-negative real values. Figure 1 shows the directed acyclic graph (DAG) representing the generative process of the LDA, including the relationships among the variables and/or parameters.

Let $\mathbf{X}$ be a random vector having support $\mathcal{S}_P$ and $\mathbf{X} \sim \text{Dir}(\mathbf{a})$; then its probability density function (p.d.f.) is expressed as

$$f_{Dir}(\mathbf{x}|\mathbf{a}) = \frac{\Gamma(a^+)}{\prod_{p=1}^P \Gamma(a_p)} \prod_{p=1}^P x_p^{a_p - 1}, \tag{1}$$

where $\mathbf{x} \in \mathcal{S}_P$, $\mathbf{a} = (a_1, \ldots, a_P)^\intercal \in \mathbb{R}_+^P$, $a^+ = \sum_{p=1}^P a_p$, and $\Gamma(\cdot)$ is the Gamma function. When all or some of the a_p's are smaller than 1, then (1) is going to be sparse (i.e., large unbounded density assigned toward all or some vertices of the simplex). Special cases of the Dirichlet include the symmetric Dirichlet (if $a_p = a$ for $p = 1, \ldots, P$) and the uniform distribution on the simplex (namely, the symmetric Dirichlet distribution with $a = 1$). The expected vector of $\mathbf{X}$ has elements $\mathbb{E}[X_p] = a_p/a^+$, whereas its covariance matrix has elements

$$\text{Var}(X_p) = \frac{\mathbb{E}[X_p](1 - \mathbb{E}[X_p])}{a^+ + 1} \quad \text{and} \quad \text{Cov}(X_p, X_{p'}) = -\frac{\mathbb{E}[X_p]\,\mathbb{E}[X_{p'}]}{a^+ + 1}, \tag{2}$$

with $p, p' \in \{1, \ldots, P\}$ and $p \neq p'$.

The Dirichlet distribution exhibits several properties that make it suitable for modeling compositional data (Aitchison 2003; Ongaro and Migliorati 2013), among which is its conjugacy with respect to the multinomial (and, by extension,

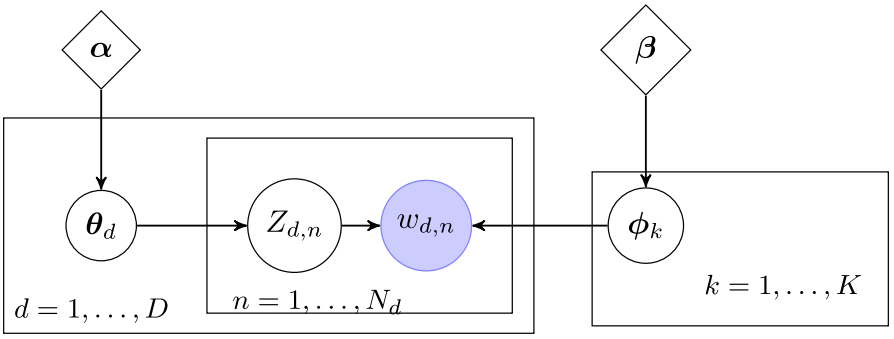

Fig. 1 DAG describing LDA model. The unobserved variables are drawn as empty circles, whereas the observed variables are filled with blue color. Rhombuses represent hyperparameters. The outer rectangle represents documents, while the inner one represents the words within a document. The rectangle containing ϕ_k represents topics

the categorical) distribution. However, the Dirichlet imposes a rigid structure of (in) dependencies. In fact, the Dirichlet distribution satisfies all the standard notions of independence defined on the simplex (e.g., neutralities, subcompositional independence, compositional invariance, etc.), which may be an unrealistic assumption in practical applications. Moreover, as shown in (2), once the expected vector $\mathbb{E}[\mathbf{X}]$ is fixed, all variances and covariances are determined by a single scalar parameter a^+. Furthermore, covariances are always negative and proportional to the product of the corresponding expected values.

In the context of LDA, imposing a Dirichlet distribution over each $\boldsymbol{\theta}_d$ leads to *near-independence* among topics, thereby excluding scenarios in which certain topics may be positively associated, which is a situation that can arise even though the simplex geometry naturally induces negative dependencies. Due to the rigidity of the Dirichlet distribution, several authors have proposed latent topic models that modify the distribution imposed on the $\boldsymbol{\theta}_d$ parameters (Blei and Lafferty 2007; LeBlanc et al. 2023).

A particularly appealing direction involves adopting distributions that allow for positive correlations among topics. One such proposal is the extended flexible Dirichlet (EFD, Ongaro et al. (2020)), a distribution defined on the simplex that includes the Dirichlet as a special case. Let $\boldsymbol{\theta}_d \sim \text{EFD}(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$, then its p.d.f. can be expressed as

$$f_{EFD}(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \frac{1}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \sum_{j=1}^{K} p_j \frac{\Gamma(\alpha_j)\Gamma(\alpha^+ + \tau_j)}{\Gamma(\alpha_j + \tau_j)} \theta_j^{\tau_j}, \quad (3)$$

where $\boldsymbol{\theta} \in \mathcal{S}_K, \boldsymbol{\alpha} \in \mathbb{R}_+^K, \boldsymbol{\tau} \in \mathbb{R}_+^K$ are two vectors with positive entries, and $\mathbf{p} \in \mathcal{S}_K$. A key feature of the EFD distribution is that it can be expressed as a finite structured (i.e., non-generic) mixture with Dirichlet components, as its p.d.f. can be rewritten as

$$f_{EFD}(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{k=1}^{K} p_k f_{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha} + \tau_k \mathbf{e}_k),$$

where $\mathbf{e}_k$ denotes the standard basis vector with 1 in the k-th position and 0 elsewhere. In this expression, the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ jointly define the component-specific Dirichlet parameters, while $\mathbf{p}$ represents the vector of mixing weights. Note that the EFD includes the Dirichlet as a special case when $\tau_k = \tau = 1$ and $p_k = \alpha_k/\alpha^+$. One of the most appealing features of the EFD mixture structure is its ability to produce densities that remain bounded while placing substantial mass near (some of) the simplex vertices.

Thanks to its additional parameters, the EFD provides greater flexibility in both its p.d.f. and covariance matrix. In particular, it introduces $2K - 1$ additional parameters compared to a standard Dirichlet distribution (namely, the K elements of $\boldsymbol{\tau}$ and $K - 1$ free elements of $\mathbf{p}$). This richer parameterization makes it possible to specify more general first two-order moments, which are given by

$$\mathbb{E}[\theta_k] = \alpha_k k_1 + p_k \frac{\tau_k}{\alpha^+ + \tau_k}, \quad (4)$$

$$\begin{aligned}\text{Var}\,(\theta_k) =&\alpha_k^2(k_2 - k_1^2) + \frac{p_k\tau_k(2\alpha_k + \tau_k + 1)}{(\alpha^+ + \tau_k)(\alpha^+ + \tau_k + 1)} + \\ &+\alpha_k k_2 - \frac{p_k^2\tau_k^2}{(\alpha^+ + \tau_k)^2} - k_1\frac{2\alpha_k p_k\tau_k}{\alpha^+ + \tau_k},\end{aligned} \tag{5}$$

and

$$\begin{aligned}\text{Cov}\,(\theta_k, \theta_{k'}) =&\alpha_k\alpha_{k'}(k_2 - k_1^2) + \frac{p_k p_{k'}\tau_k\tau_{k'}}{(\alpha^+ + \tau_k)(\alpha^+ + \tau_{k'})} + \\ &+\frac{\alpha_k p_{k'}\tau_{k'}}{\alpha^+ + \tau_{k'}}\left(\frac{1}{\alpha^+ + \tau_{k'} + 1} - k_1\right) + \\ &+\frac{\alpha_{k'} p_k\tau_k}{\alpha^+ + \tau_k}\left(\frac{1}{\alpha^+ + \tau_k + 1} - k_1\right),\end{aligned} \tag{6}$$

where $k_1 = \sum_{j=1}^{K}\frac{p_j}{\alpha^+ + \tau_j}$ and $k_2 = \sum_{j=1}^{K}\frac{p_j}{(\alpha^+ + \tau_j)(\alpha^+ + \tau_j + 1)}$.

Ongaro et al. (2020) showed that (even large) positive covariances under the EFD distribution can be obtained by inducing a large variability in the elements of $\boldsymbol{\tau}$. A more practical way to induce large variability in the elements of $\boldsymbol{\tau}$ is to rely on an alternative parameterization. In this regard, Ascari et al. (2024) proposed an alternative formulation of the EFD distribution that explicitly includes the marginal barycenter $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\theta}]$ together with the vector $\tilde{\mathbf{w}}$, whose j-th component is defined as

$$\tilde{w}_j = \frac{\tau_j}{\alpha^+ + \tau_j}\cdot\frac{1}{\min\left(1, \frac{\mu_j}{p_j}\right)}, \tag{7}$$

where the second factor normalizes the parameter space, ensuring that $\tilde{w}_j \in (0, 1)$. Equation (7) shows that, as τ_j increases, $\tilde{w}_j$ approaches 1, thereby preserving the interpretation of τ_j in the new parameterization. Large variability among the τ_j's can be achieved by choosing the $\tilde{w}_j$'s to be approximately equidistant. This parameterization is going to be useful when dealing with hyperparameters' initialization in Sect. 5. Section S1.1 of the SM shows the different parameterizations of the EFD distribution and the relationships linking them, whereas Section S1.2 deepens the interpretation of the role of the hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\tau}$, and $\mathbf{p}$. Unlike a generic (i.e., unconstrained) mixture of Dirichlet distributions, the EFD offers several advantages. First, the links between component-specific parameters ensure identifiability of the model, a property that generally cannot be guaranteed for unconstrained mixtures. Second, for a generic Dirichlet mixture, the first- and second-order moments must be derived for each specific parameter configuration, with no guarantee that positive dependence among components can arise or that it can be induced through a simple and interpretable mechanism such as the one described above. Finally, it is worth noting that the EFD preserves some key properties of the Dirichlet distribution, including conjugacy with respect to multinomial and categorical likelihoods.

Proposition 1 *Let* $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_D)^\intercal$ *be a sample with i.i.d. elements generated from* $\mathbf{X}_d|\boldsymbol{\theta} \sim$ *Multinomial*$(N_d, \boldsymbol{\theta})$, $d = 1, \ldots, D$. *Then, the EFD family is conjugated to the multinomial distribution, namely if* $\boldsymbol{\theta} \sim$ *EFD*$(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$, *then* $\boldsymbol{\theta}|\mathbf{x} \sim$ *EFD*$(\boldsymbol{\alpha}^*, \boldsymbol{\tau}, \mathbf{p}^*/p_+^*)$, *where* $\boldsymbol{\alpha}^* = \boldsymbol{\alpha} + \mathbf{x}^+$, $\mathbf{p}^*/p_+^* \in \mathcal{S}_K$ *has k-th element* $p_k^* = p_k \frac{(\alpha_k + \tau_k)^{[x_k^+]}}{(\alpha^+ + \tau_k)^{[N]}(\alpha_k)^{[x_k^+]}}$, $N = \sum_{d=1}^{D} N_d$, $p_+^* = \sum_{k=1}^{K} p_k^*$, $\mathbf{x}^+ = (x_1^+, \ldots, x_K^+)^\intercal$, $x_k^+ = \sum_{d=1}^{D} x_{k,d}$, *and* $a^{[m]} = a(a+1)\ldots(a+m-1)$ *is the rising factorial function, with* $a^{[0]} = 1$.

For the proof, see Section S2 of the SM.

Thus, we define the EFLDA by assuming that

$$\boldsymbol{\theta}_d \sim \mathrm{EFD}(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) \quad \text{and} \quad \boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\beta}).$$

The EFLDA contains both the standard LDA and the flexible LDA (FLDA, Giampino et al. (2025)) as special cases. Specifically, the LDA is recovered by imposing the previously described parameter structure for which the EFD reduces to the Dirichlet distribution, whereas the FLDA is obtained when $\tau_k = \tau$, $k = 1, \ldots, K$. The FLDA has emerged as an interesting mixed-membership model, serving as the counterpart of latent topic models in microbiome data analysis (Deek et al. 2021; LeBlanc et al. 2023). However, it cannot accommodate positive covariances between *communities* (i.e., the analogue of topics). While FLDA allows for more general covariance structures than LDA, it still only permits negative covariances.

Figure 2 illustrates the DAG of the EFLDA generative process. Unlike the DAG in Fig. 1, it includes a set of latent variables Z_d^*, which represent the EFD mixture-component label for document d. Specifically, $Z_d^* = j$ indicates that the vector of topic proportion $\boldsymbol{\theta}_d$ is drawn from the j-th mixture component. Consequently, in generating the topic-proportion vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D$ (step (ii) of the previous corpus-generation process), one first samples a mixture-component label z_d^* from $Z_d^* \sim \mathrm{Categorical}(\mathbf{p})$ and then generates $\boldsymbol{\theta}_d$ from $\boldsymbol{\theta}_d \sim Z_d^* = j \sim \mathrm{Dir}(\boldsymbol{\alpha} + \tau_j \mathbf{e}_j)$.

By selecting a mixture-based distribution for the $\boldsymbol{\theta}_d$'s, as in the EFLDA model, we introduce an additional layer of structure into the latent topic model. This choice allows for the identification of clusters of topics, which we refer to as *sub-topics*. A sub-topic can be interpreted as a specific combination of topics, characterized by a distinctive balance among them. In particular, EFLDA controls the degree of overlap among sub-topics via the vector $\boldsymbol{\tau}$: the larger τ_j, the more distinct the j-th sub-topic becomes, eventually collapsing into a single topic as τ_j further increases.

Sub-topics can enrich corpus analysis by offering a deeper representation of textual documents in terms of topic proportions and by supporting more impactful communication. By way of example, let us consider the ternary diagram in Fig. 3, representing the joint p.d.f. of a particular EFD distribution with well separated mixture components, imposed on three latent topics, namely *Mathematics* (θ_1), *Economics* (θ_2), and *Politics* (θ_3).

In this representation, it is clear that the EFD assigns high density to three non-overlapping subsets of the simplex, each corresponding to the neighborhood of a distinct mode. Specifically, the upper region can be labeled *Political Economy*, as it lies

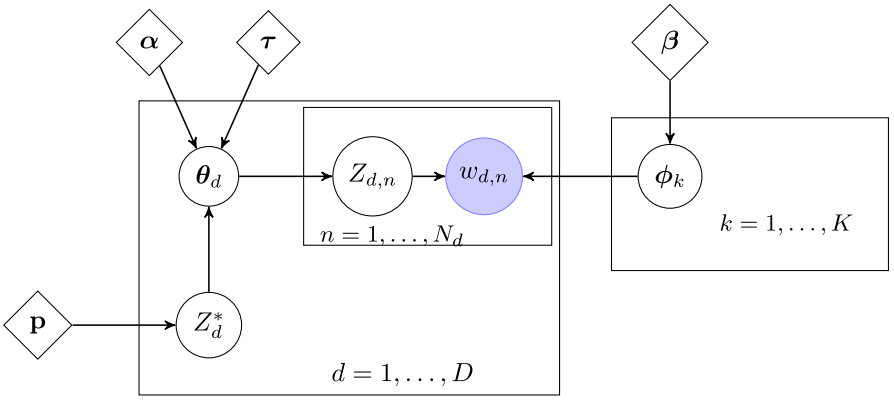

Fig. 2 DAG describing EFLDA model. The unobserved variables are drawn as empty circles, whereas the observed variables are filled with blue color. Rhombuses represent hyperparameters. The outer rectangle represents documents, while the inner one represents the words within a document. The rectangle containing ϕ_k represents topics

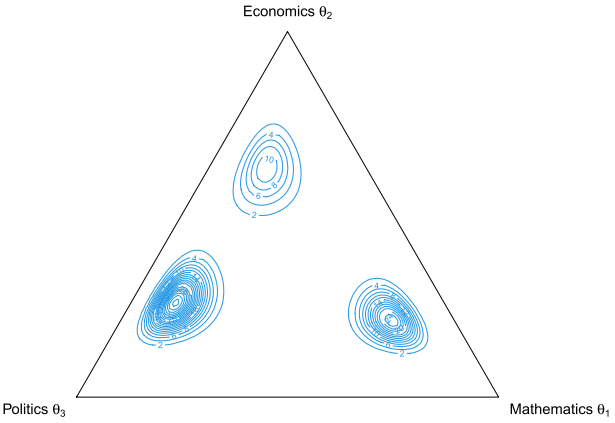

Fig.3 Ternary diagram reporting the p.d.f. of an EFD with parameters $\boldsymbol{\alpha} = (8, 18, 13)^\intercal$, $\boldsymbol{\tau} = (45, 15, 30)^\intercal$, and $\mathbf{p} = (0.3, 0.2, 0.5)^\intercal$

closer to the Economy vertex and its second-closest vertex corresponds to Politics. Similarly, the lower-left and lower-right subsets can be labeled *Economic Policy* and *Theoretical Economics*, respectively.

It is worth noting that the Dirichlet distribution (and, consequently, LDA) cannot generate a p.d.f. that concentrates density on specific regions of the simplex while assigning almost zero density elsewhere. To ensure non-negligible density across different regions, one must instead use a Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (1, 1, \ldots, 1)^\intercal$, which corresponds to the uniform distribution on the simplex.

3 Collapsed Gibbs sampling

Fitting a latent topic model, such as the LDA and the EFLDA, involves estimating two distinct sets of parameters, namely the set of document-specific topic proportions $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$ and the set of topic-specific word distributions $\Phi = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$. In this perspective, it is important to distinguish between two types of sample sizes. The first, and most intuitive, is the number D of documents in the corpus: a larger D typically improves the estimation of each $\boldsymbol{\phi}_k$. The second is the document length, denoted by N_d, which represents the number of words in document d and influences the estimation of the corresponding $\boldsymbol{\theta}_d$, $d = 1, \ldots, D$.

It is well-known that estimation of the elements in Θ and Φ via classical tools, such as maximum likelihood or posterior-based inference, is challenging, mainly due to the presence of the latent topic labels (Blei et al. 2003; Griffiths and Steyvers 2004; Giampino et al. 2025).

Thus, to estimate the parameters in Θ and Φ, we consider the CGS algorithm (Liu 1994). In this version of the standard Gibbs sampler (GS), the goal is to generate samples from the joint posterior $f(\mathbf{Z}, \Theta, \Phi | \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\beta})$, where $\mathbf{Z} = \{Z_{d,1}, \ldots, Z_{d,N_d}\}_{d=1}^{D}$ denotes the collection of topic labels and $\mathcal{C}$ is the observed corpus through the DTM (i.e., the only observed data). However, this posterior distribution is difficult to sample from, even in simpler models such as the LDA (Griffiths and Steyvers 2004). Therefore, in the CGS we marginalize some parameters out from the joint posterior and subsequently sample from the full conditionals of the remaining variables. The marginalized parameters can then be estimated by exploiting some conjugacy properties.

In our framework, we marginalize the parameters (Θ, Φ) out, so that the GS can be sketched for sampling $\mathbf{Z}$ from $f(\mathbf{Z} | \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\beta})$. Estimates for generic $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ can be obtained by resorting to conjugacy properties of Dirichlet and/or EFD priors with respect to multinomial/categorical distributions.

The full conditionals of interest are the probabilities that $Z_{d,n} = k\,(k \in \{1, \ldots, K\})$, given all other topic labels $\mathbf{z}_{-(d,n)}$, the observed data, and the hyperparameters. To compute these conditionals under the EFLDA model, it is necessary to introduce some key quantities related to the counts. Let $c_{d,k,v} = \sum_{n=1}^{N_d} \mathbb{I}(z_{d,n} = k, w_{d,n} = v)$ be the number of times word v is assigned to topic k in the d-th document of the corpus. Then, by summing across the pertinent indices, we define the number $c_{d,k,\cdot}$ of words assigned to topic k in document d, the number $c_{\cdot,k,v}$ of times that word v is assigned to topic k across documents, and the number $c_{\cdot,k,\cdot}$ of total words assigned to the k-th topic. Lastly, notation c^{-} represents the same count excluding the n-th word of document d (i.e., the one we are computing the full conditional for).

Given these quantities, the full conditional for $Z_{d,n}$ is

$$\mathbb{P}(Z_{d,n} = k|\mathbf{z}_{-(d,n)}, \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\beta}) \propto \frac{\left(\alpha_k + c^-_{d,k,\cdot}\right)\left(\beta_{w_{d,n}} + c^-_{\cdot,k,w_{d,n}}\right)}{\left(\beta^+ + c^-_{\cdot,k,\cdot}\right)} \cdot \left\{\sum_{j=1}^{K} \frac{p^*_{d,j}}{\left(\alpha^+ + \tau_j + N^-_d\right)} + \frac{p^*_{d,k}}{\left(\alpha^+ + \tau_k + N^-_d\right)}\left(\frac{\tau_k}{\alpha_k + c^-_{d,k,\cdot}}\right)\right\}, \tag{8}$$

where

$$p^*_{d,k} = p_k \frac{(\alpha_k + \tau_k)^{[c_{d,k,\cdot}]}}{(\alpha_k)^{[c_{d,k,\cdot}]}(\alpha^+ + \tau_k)^{[N^-_d]}}. \tag{9}$$

By setting $\tau_k = 1$ and $p_k = \alpha_k/\alpha^+$ in Equation (8), we recover the LDA's full conditional derived in Griffiths and Steyvers (2004). Interestingly, replacing the Dirichlet prior on each $\boldsymbol{\theta}_d$ with the EFD distribution modifies the full conditional only through the inclusion of an additional factor (i.e., the term in curly brackets), which depends on the parameters characterizing the mixture components, namely $\boldsymbol{\tau}$ and $\mathbf{p}$. On the one hand, each $p^*_{d,k}$ can be interpreted as an updated (unnormalized) mixture weight that also accounts for the observed counts $c_{d,k,\cdot}$ and cluster distances. On the other hand, τ_k regulates how strongly the k-th topic is weighted relative to the others in determining the full conditional distribution of the associated label.

Full conditionals in Equation (8) represent the kernel of a probability mass function, so generating MCMC samples for $\mathbf{z}$, namely $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(B)}$, is straightforward. Once the MCMC sample $\mathbf{z}^{(b)}, b = 1, \ldots, B$, has been obtained, we need a rule to estimate the parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$. A natural approach, which is consistent with Griffiths and Steyvers (2004), is to exploit the conjugacy properties of the Dirichlet and EFD distributions introduced in Sect. 2. In fact, it can be shown that the posterior distributions of $\boldsymbol{\theta}_d$, conditional on the topic assignments, is

$$\boldsymbol{\theta}_d|\mathbf{z}, \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p} \sim \text{EFD}(\boldsymbol{\alpha} + \mathbf{c}_d, \boldsymbol{\tau}, \mathbf{p}^*_d/p^{*+}_d), \tag{10}$$

where $\mathbf{c}_d = (c_{d,1,\cdot}, \ldots, c_{d,K,\cdot})^\intercal$, $\mathbf{p}^*_d = \left(p^*_{d,1}, \ldots, p^*_{d,K}\right)^\intercal$, $p^{*+}_d = \sum_{k=1}^K p^*_{d,k}$, and $p^*_{d,k}$ is given by Equation (9). Similarly, the posterior distribution of $\boldsymbol{\phi}_k$, given the topic assignments, is

$$\boldsymbol{\phi}_k|\mathbf{z}, \mathcal{C}, \boldsymbol{\beta} \sim \text{Dir}(\boldsymbol{\beta} + \mathbf{d}_k), \tag{11}$$

where $\mathbf{d}_k = (c_{\cdot,k,1}, \ldots, c_{\cdot,k,V})^\intercal$. Note that the posterior distribution of $\boldsymbol{\phi}_k$ has the same form as in the LDA model, since both models share the same Dirichlet prior on $\boldsymbol{\phi}_k$. Nevertheless, because the models infer different topics, their resulting posterior distributions (i.e., their updated hyperparameters) will inevitably differ.

Given these results, a reasonable choice is to estimate $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ with their posterior means based on $\mathbf{z}^{(b)}$, so that

$$\hat{\theta}_{d,k}^{(b)} = \left(\alpha_k + c_{d,k,\cdot}^{(b)}\right) k_{1,d}^{(b)} + \left(\frac{p_{d,k}^{*(b)}}{p_d^{*+(b)}}\right) \frac{\tau_k}{\alpha^+ + \tau_k + N_d}, \tag{12}$$

where

$$k_{1,d}^{(b)} = \sum_{j=1}^{K} \frac{p_{d,j}^{*(b)}}{p_d^{*+(b)}} \frac{1}{\alpha^+ + \tau_j + N_d},$$

and

$$\hat{\boldsymbol{\phi}}_k^{(b)} = \frac{\boldsymbol{\beta} + \mathbf{d}_k^{(b)}}{\beta^+ + c_{\cdot,k,\cdot}^{(b)}}. \tag{13}$$

Thus, the b-th step of the CGS consists in generating $\mathbf{z}^{(b)}$ and then use it for computing $\boldsymbol{\theta}_d^{(b)}$ and $\boldsymbol{\phi}_k^{(b)}$.

The choice of the number of topics K is a central challenge in topic models, since it must be specified by the user. A common strategy is to fit the model for multiple candidate values of K and subsequently evaluate performance using a suitable metric. Ideally, when K is sufficiently large, the assignment of words to topics should remain stable, with superfluous topics receiving negligible probability mass (Wallach et al. 2009). However, in practical applications, the robustness of model inference to misspecified values of K remains largely unclear. Selecting K that is too small or excessively large often distorts the inferred structure, compromising both interpretability and topic structure (Wallach et al. 2009; Subeno et al. 2018). A popular metric for guiding the choice of K is the perplexity (Blei et al. 2003).

Let $N = \sum_{d=1}^{D} N_d$ be the total number of word forming the DTM; then, the log-likelihood of document d is defined as

$$\log \mathbb{P}(d) = \sum_{w \in d} \log \mathbb{P}(w \mid d) = \sum_{w \in d} \log \sum_{k=1}^{K} \phi_{k,w} \theta_{d,k}, \tag{14}$$

with the summation $\sum_{w \in d}$ spanning all the observed words in document d. Specifically, given a corpus $\mathcal{C} = \{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$, the perplexity is defined as the exponentiated negative average log-likelihood of the dataset, so that lower values correspond to better predictive performance, as

$$\text{Perplexity}(\mathcal{C}) = \exp\left\{-\frac{\sum_d \log \mathbb{P}(d)}{N}\right\}. \tag{15}$$

An empirical estimate of perplexity can be obtained by plugging the posterior parameter estimates from the MCMC output into Equation (14) and then computing

the average values across MCMC iterations, thus obtaining an MCMC perplexity estimate.

For new data, namely previously unseen documents, we consider a held-out test set consisting of D' new documents for which the topic assignments are unknown. In this setting, it is necessary to estimate the topic proportions $\hat{\boldsymbol{\theta}}_{D+1}, \ldots, \hat{\boldsymbol{\theta}}_{D+D'}$, which can be computationally demanding. Following the strategy proposed by Yao et al. (2009), we update the topic assignments only for the new documents while keeping those from the training set fixed. This allows the topic proportions for the test set to be inferred using the MCMC procedure, substantially reducing computational cost without compromising inferential accuracy.

All probabilistic models (EFLDA, its simplified version FLDA, and the standard LDA model), as well as functions to deal with their outputs, were implemented in C++ and wrapped in R within the `EFLDA` package, which is fully available in the GitHub repository at https://github.com/AliceGiampino/EFLDA.

4 Discovering sub-topics

As stated in Sect. 2, introducing a finite mixture distribution for the $\boldsymbol{\theta}_d$'s allows us to define sub-topics. More specifically, sub-topics correspond to the mixture components which, under the EFD distribution, are Dirichlet-distributed.

Once having the MCMC sample $\mathbf{z}^{(b)}$, $b = 1, \ldots, B$, it is possible to estimate the posterior probability $\hat{q}_{d,k}^{(b)}$ that a given document d, represented by its vector of topic proportion $\boldsymbol{\theta}_d$, arises from the k-th Dirichlet component:

$$\hat{q}_{d,k}^{(b)} = \frac{p_k^{*(b)} f_{Dir}\left(\hat{\boldsymbol{\theta}}^{(b)}; \boldsymbol{\alpha} + \mathbf{c}_d^{(b)} + \tau_k \mathbf{e}_k\right)}{\sum_{k'=1}^{K} p_{k'}^{*(b)} f_{Dir}\left(\hat{\boldsymbol{\theta}}^{(b)}; \boldsymbol{\alpha} + \mathbf{c}_d^{(b)} + \tau_{k'} \mathbf{e}_{k'}\right)}. \tag{16}$$

These posterior probabilities can be used for model-based clustering, assigning each textual document to one of the sub-topics (i.e., mixture components). More specifically, document d is allocated to sub-topic j if and only if $\hat{q}_{d,j}^{(b)} > \hat{q}_{d,k}^{(b)}$ for all $k \neq j$, that is, if the posterior probability that $\hat{\boldsymbol{\theta}}_d$ originated from mixture component j is the largest.

In order to make the concept of sub-topics relevant, we must assign a label to each sub-topic, as done in Fig. 3 (and analogously to the labeling of actual topics). Each sub-topic can, in principle, be associated with any of the $\mathcal{D}_{K,2} = K!/(K-2)!$ possible ordered labels obtained by pairing two topics. To associate labels to sub-topics, suppose that $\hat{\boldsymbol{\theta}}_d^{(b)}$ has been assigned to sub-topic j. We then need to assign a label to the component $\mathrm{Dir}(\boldsymbol{\alpha} + \mathbf{c}_d^{(b)} + \tau_j \mathbf{e}_j)$. Our proposal is to approximate the distance between this Dirichlet distribution and each vertex of the K-part simplex. Specifi-

cally, we rely on an additional Monte Carlo (MC) step based on T iterations and compute

$$d(j,k)^{(b)} = \frac{1}{T}\sum_{t=1}^{T} d_{SKL}\left(\mathbf{x}_{j,b}^{(t)}, \mathbf{e}_k\right), \quad k = 1,\ldots,K,$$

where the $\mathbf{x}_{j,b}^{(t)}$ are i.i.d. draws from $\text{Dir}(\boldsymbol{\alpha} + \mathbf{c}_d^{(b)} + \tau_j \mathbf{e}_j)$, and $d_{SKL}(\cdot,\cdot)$ denotes the symmetric Kullback–Leibler (SKL) divergence (Kullback and Leibler 1951). More formally, if $\mathbf{x}$ and $\mathbf{y}$ are two compositional vectors defined on the same simplex $\mathcal{S}_P$, then

$$d_{SKL}(\mathbf{x},\mathbf{y}) = d_{KL}(\mathbf{x},\mathbf{y}) + d_{KL}(\mathbf{y},\mathbf{x}), \text{ where } d_{KL}(\mathbf{x},\mathbf{y}) = \sum_{p=1}^{P} x_p \log\left(\frac{x_p}{y_p}\right). \quad (17)$$

It is important to note that the SKL divergence, like many other metrics for compositional data, cannot be computed at the edges or vertices of the simplex due to zeros appearing in the logarithm and/or in the denominator of the ratio. Therefore, instead of using the exact simplex vertex $\mathbf{e}_k$, we consider a perturbed point $\mathbf{e}_k^*$ having $1-\varepsilon$ in position k and $\varepsilon/(K-1)$ in all other positions ($\varepsilon = 0.01$).

By selecting the two (ordered) values of k that yield the smallest $d(j,k)^{(b)}$, we can assign a label to sub-topic j for document d in the b-th MCMC iteration. In the example in Fig. 3, labeling sub-topic j, corresponding to the lower-left mixture component, requires computing $d(j,1)^{(b)}$, $d(j,2)^{(b)}$, and $d(j,3)^{(b)}$. Suppose the smallest value is $d(j,3)^{(b)}$ and the second smallest is $d(j,2)^{(b)}$; this yields the sub-topic label "3-2". Under the assumption that Topic 3 represents Politics and Topic 2 represents Economics, this sub-topic can be labeled *"Economic Policy"*.

5 Simulation study for initialization strategies

Latent topic models, including LDA and its extensions, require the specification of several hyperparameters to properly fit the model. The classical LDA model involves choosing values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, whereas the EFLDA additionally requires the specification of $\boldsymbol{\tau}$ and $\mathbf{p}$. In the LDA framework, heuristic guidelines for hyperparameter initialization are commonly adopted by the research community (Griffiths and Steyvers 2004; Deek et al. 2021; Giampino et al. 2025). In particular, non-informative settings often assume symmetric Dirichlet priors, such that $\alpha_k = \alpha$ for $k = 1,\ldots,K$ and $\beta_v = \beta$ for $v = 1,\ldots,V$. Typical choices for these parameters include $\alpha \in \{1, 1/K, 50/K\}$ and $\beta \in \{0.01, 0.02, 1\}$.

To determine the most appropriate initialization strategy for the EFLDA model, we conducted a sensitivity analysis aimed at evaluating different initialization strategies for the hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\tau}$, and $\mathbf{p}$ in terms of hyperparameters recovery. Six different simulation settings were considered, while keeping the dataset dimensions constant across all scenarios. Corpora consisted of a vocabulary of size $V = 150$ and $D = 200$ documents, each having an average length $N_d \sim \text{Pois}(50)$. For each

scenario, we generated $R = 100$ synthetic corpora and assessed the performance of different initialization methods by measuring how precisely they recovered the true hyperparameters, quantified through distance metrics. In each simulating scenario, we set the true value of β equal to the V-dimensional unit vector (i.e., $\boldsymbol{\beta}_{\text{true}} = (1, 1, \ldots, 1)^\intercal$).

We considered six scenarios, each characterized by a different combination of the number of topics K and the correlation structure among the elements of each $\boldsymbol{\theta}_d$. Specifically, we generated corpora from an EFLDA model with $K \in \{3, 5, 10\}$. For each value of K, we defined two additional scenarios: the first characterized by positive correlations among topic proportions, and the second by a general correlation matrix with only negative values.

To generate corpora with positive correlations, we set the true values of the hyperparameters using the alternative parameterization described in Sect. 2. Specifically, instead of selecting the elements of $\boldsymbol{\tau}$ with large variability, we choose the elements of $\tilde{\mathbf{w}} \in (0, 1)^K$ to be almost equally spaced. Thus, rather than fixing the true values of $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$, and $\mathbf{p}$, we set the values of $\boldsymbol{\mu}$, $\tilde{\mathbf{w}}$, and α^+, and then map them back to the original parameter space (explicit expressions can be found in Section S1.1 of the SM). The true parameter values (for both parameterizations) are reported in Section S4 of the SM.

Regarding the initialization methods, we consider three distinct schemes. Each strategy involves fitting a preliminary EFLDA model with non-informative or weakly informative hyperparameters, and then using the corresponding posterior means $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_D$ to derive estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$, and $\mathbf{p}$ as described below, which can subsequently be used to fit the final EFLDA model.

All three initialization schemes share the choice of $\mathbf{p} = (1/K, \ldots, 1/K)^\intercal$, but differ in the assumed mixture structure of the underlying EFD distribution. Specifically, we consider the following initialization methods:

- Method 1 (*EFLDA with positive correlations*). This initialization aims to facilitate the exploration and detection of positive dependencies. For this reason, we initialize $\tilde{\mathbf{w}}$ and $\boldsymbol{\mu}$ instead of $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$. The vector $\tilde{\mathbf{w}}$ is defined as an equally spaced sequence between 0.1 and 0.9, whereas $\boldsymbol{\mu}$ is set to be non-informative, namely $\boldsymbol{\mu} = (1/K, \ldots, 1/K)^\intercal$, and $\alpha^+ = 100$. The initial values for $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ are then obtained by mapping $\boldsymbol{\mu}$ and $\tilde{\mathbf{w}}$ back to the original space.
- Method 2 (*EFLDA with a generic dependence structure*). The goal of this initialization is to specify an EFLDA model that is as non-informative as possible while still preserving the mixture structure typical of the EFD distribution. To this end, we set $\alpha_k = \alpha = 50/K$ for $k = 1, \ldots, K$. The remaining hyperparameter to define is $\boldsymbol{\tau}$. To retain the key feature of the EFD prior (i.e., distinct values of τ_k for at least one k), we assign arbitrary values to the components of $\boldsymbol{\tau}$. In this case, there is no any principled way to define non-informative values for $\boldsymbol{\tau}$. Therefore, we sampled the elements τ_k with replacement from the set $\{10, 20, 30, 50\}$.
- Method 3 (*FLDA*). This initialization relies on a weakly informative FLDA model. Specifically, we set $\alpha_k = \alpha = 50/K$ and $\tau_k = \tau$ for $k = 1, \ldots, K$. While the choice of τ may be guided by prior knowledge of the topic mixture structure, in this case we propose a heuristic rule: $\tau = 1.5 \cdot \alpha^+$. This choice ensures a non-

negligible separation between the Dirichlet components in the EFD prior.

In cases where some settings are coherent with the corpus generating mechanism (e.g., $\tilde{\mathbf{w}}$ with almost equally spaced elements), the true parameter values were chosen to differ from the initialization scheme, so as not to favor any particular method.

To compare the different initialization schemes, we fitted three EFLDA models to each generated corpus, using the initialization strategies described above. The resulting estimates of each $\boldsymbol{\theta}_d$ were then employed to estimate the parameters of an EFD$(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$ via a Bayesian estimation procedure implemented in the Stan software (Ascari et al. 2024; Stan Development Team 2022). The obtained estimates were subsequently used as proposed values for the hyperparameters and compared with the true hyperparameters that generated the corpora. Comparisons were made by using the Euclidean distance for $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$, and $\tilde{\mathbf{w}}$, and the d_{SKL} in Equation (17) for $\mathbf{p}$ and $\boldsymbol{\mu}$.

Since the Bayesian estimation procedure may return the elements of $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$, and $\mathbf{p}$ in an arbitrary order, identifying the correct permutation is essential for a meaningful comparison between the proposed and true hyperparameters. To do so, we evaluated all possible permutations of $\boldsymbol{\tau}$ and selected the one minimizing the ℓ_2 distance between the true $\boldsymbol{\tau}$ vector and its estimated counterpart. The optimal permutation was then consistently applied to all other parameters. After obtaining the permuted versions of the estimated quantities, we computed the metrics in (17).

Figures S5–S10 in the SM, of which Fig. 4 in the SM represents an illustrative re-arrangement for the fifth scenario, summarize the distances between the true and proposed hyperparameters across the six considered scenarios. Although no method appears to be globally optimal, the initialization based on Method 3, namely the FLDA-based approach, yields the most favorable bias-variance trade-off (i.e., narrower boxes centered at smaller values). Consequently, this initialization scheme is the one we recommend when no prior information on the hyperparameters is available.

We also compared the three methods in terms of their ability to handle positive correlations and their perplexity (Eq. 14). On the one hand, Table S7 in the SM reports the average proportion of positive pairwise correlations identified by each initialization method across B MCMC iterations. Specifically, given the posterior chains for the initializations of $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$, and $\mathbf{p}$ as described above, we compute the theoretical correlation matrix $\mathbf{R}^{(b)}$ of an EFD$(\boldsymbol{\alpha}^{(b)}, \boldsymbol{\tau}^{(b)}, \mathbf{p}^{(b)})$. Let $r_+^{(b)}$ denote the number of positive off-diagonal elements in the upper triangular portion of $\mathbf{R}^{(b)}$. The table displays the average proportion:

$$\frac{1}{B}\sum_{b=1}^{B}\frac{r_+^{(b)}}{K(K-1)/2}, \tag{18}$$

where $K(K-1)/2$ is the total number of unique pairwise correlations. All initialization schemes are able to identify some positive dependencies when the corpora are generated from models that include positively correlated topics. The main difference emerges in the scenarios without positive dependencies: in these cases, only the

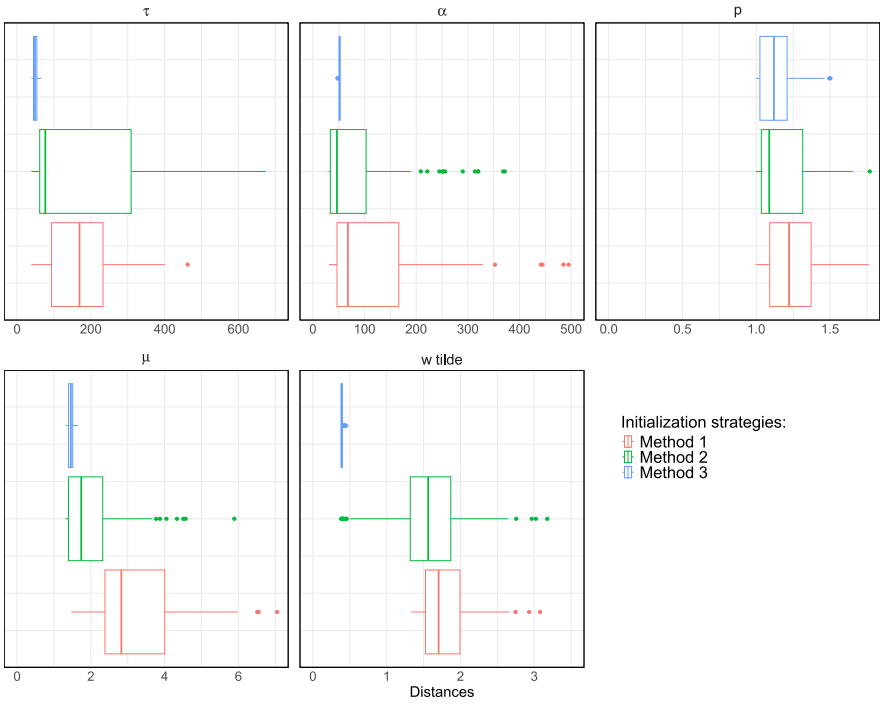

Fig. 4 Boxplots of the distances between the true and proposed hyperparameters in the case of $K = 5$ after the labels permutation (Scenario E)

FLDA-based initialization exhibits sufficient robustness to avoid introducing spurious positive correlations in the prior.

On the other hand, Figure S11 reports the perplexity for the models with the different initialization methods for the considered scenarios. For small values of K, Method 3 tends to perform slightly better than the others. However, as K increases, the differences between the methods progressively diminish and eventually become negligible. This suggests that, while initialization may have a modest impact in low-dimensional settings, EFLDA becomes increasingly robust to the choice of initialization as the number of topics grows.

To provide additional insight into the computational cost of the proposed CGS algorithm, we report in Figure S14 of the SM the runtime of the sampler as a function of the vocabulary size V and the number of documents D in the simulation study. The results show that the computational cost increases approximately linearly with both V and D, which is consistent with the structure of the sampler. On a standard laptop with Windows OS, 8 cores, and `R` version 4.3.3, the full procedure (including warm-up and post-processing) typically runs in the order of a few minutes for the largest simulated settings considered here. As expected, increasing either the number of documents or the vocabulary size leads to a proportional increase in runtime, while the method remains computationally manageable for moderate-scale problems. Details on the MCMC convergence are reported in Section S4.2 of the SM.

6 Applications

In the following, we consider three applications to evaluate the EFLDA model across different domains. The first and main application concerns the "Mental Health Counseling Conversations" dataset,[1] where we aim to uncover latent thematic structures and identify sub-topics related to emotional expression, coping strategies, and support-seeking behaviors. The second application focuses on microbiome data, a field in which latent topic models are often referred to as mixed-membership models. In this case, we consider the COMBO dataset (Wu et al. 2011; Giampino et al. 2025) and aim to identify latent communities (i.e., the counterpart of topics) of taxa based on their co-occurrence patterns. The third and final application involves a more standard textual domain, namely a collection of BBC Sport news,[2] with the objective of evaluating the ability of the model to recover coherent and interpretable topics in a structured corpus with well-defined thematic categories. In these applications, we consider six competing models: the probabilistic LDA, FLDA, and EFLDA, together with the non-probabilistic approaches NMF, BERTopic, and Top2Vec (a description of NMF, BERTopic, and Top2Vec is provided in Section S5 of the SM).

6.1 Mental health counseling conversations

The "Mental Health Counseling Conversations" dataset contains 3512 individual mental health counseling sessions between patients and licensed professionals. Each record consists of a question-answer pair, where several specialists independently respond to specific issues presented by a patient.

Within the 3512 pairs, duplicates were detected for both questions and answers, although their origins differ. Question duplicates primarily arise because identical issues are posed to multiple professionals, while answer duplicates are mainly due to inconsistencies introduced during data collection. After removing duplicates, the dataset comprises 995 unique questions and 2480 unique answers. While the two-variable structure allows, in principle, for examining both user concerns and professional responses, the current application concentrates exclusively on the questions for the extraction of latent themes.

Thus, questions underwent standard preprocessing, including tokenization, lowercasing, removal of punctuation, numbers, and English stopwords. Finally, English stemming was applied to standardize lexical forms. We also excluded from the analysis words occurring less than 50 times in the questions, as well as records that were too short to provide sufficient information for reliable topic modeling. This process led to a final DTM characterizing a corpus of 143 documents and 139 unique words.

In the analysis of this dataset, we considered values of $K \in \{3, \ldots, 21\}$. To select the optimal number of topics for each model, the data were randomly partitioned into a training set (80%) and a test set (20%). Models were estimated using the training set, and their performance was subsequently assessed on the test set. The DTM of the training (test) set had dimensions 114×139 (29×139).

[1] https://huggingface.co/datasets/Amod/mental_health_counseling_conversations

[2] https://www.kaggle.com/datasets/maneesh99/sports-datasetbbc

Figure 5 reports the perplexity scores for the four considered models, for both the training and test sets. By inspecting the training set results (left panel), it can be observed that the EFLDA and FLDA models exhibit comparable performance, positioned between LDA and NMF. The LDA model performs worse than all the others on the training set, while NMF shows a regularly decreasing perplexity as K increases. Given the non-probabilistic nature of NMF (i.e., it is a decomposition of the DTM), this trend indicates overfitting, since larger values of K imply greater model flexibility. The perplexity values for NMF on the test set confirm this interpretation, with a minimum value of 819.94, which is much larger than the other perplexity scores. The minimum perplexity value on the test set (right panel of Fig. 5) is achieved by the EFLDA model with $K = 5$ topics.

BERTopic and Top2Vec were fitted under different data representations, namely using both raw and preprocessed text (detailed results are reported in Section S6.1 of the SM). While these models can be applied directly to raw (i.e., the original non-processed) text and do not strictly require preprocessing, we considered both settings to evaluate the impact of text normalization on the resulting topic structure. Top2Vec identifies a single topic in both settings, although the associated keywords differ between raw and preprocessed data. In contrast, BERTopic identifies three topics in the raw data, along with an extra Topic labelled as "-1" corresponding to an outlier topic and indicating that a non-negligible portion of the documents is not assigned to any coherent topic. When applied to preprocessed data, BERTopic identifies two topics and one outlier cluster. After preprocessing, the topics extracted from BERTopic exhibit enhanced interpretability, characterized by a higher prevalence of meaningful terms aligned with mental health themes. Nevertheless, the results retain certain ambiguities, such as noticeable overlap and insufficient differentiation between topics. Furthermore, BERTopic proved to be highly sensitive to preprocessing configurations; small variations in the input representation led to substantial shifts in the structure and, possibly, the overall number of identified topics.

Therefore, being the EFLDA with $K = 5$, the model that achieves the most promising results, we re-fitted this model considering this number of latent topics on the

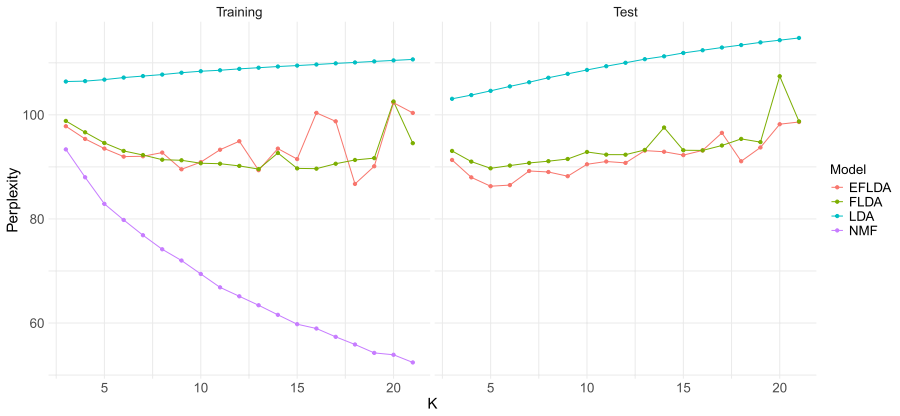

Fig. 5 Mental Health application: Perplexity value on the training (left panel) and the test set (right panel) as a function of K. In the right panel, the NMF has been removed for graphical reasons

complete corpus of 143 documents. Resulting topics are represented as word clouds in Fig. 6. Topic labels were assigned through qualitative inspection of the ten most probable words for each topic (i.e., the ten words characterized by the largest $\hat{\phi}_{k,v}$, for topic k). Topic 1 includes words such as 'don't', 'feel', 'like', 'wrong', and 'talk', which point to difficulties in expressing emotions and communicating effectively. Topic 2 is characterized by terms related to domestic life and family dynamics (e.g., 'parent', 'house', 'live', 'mom', 'boyfriend'), suggesting issues surrounding living arrangements and interpersonal tensions. Topic 3 contains emotionally salient words such as 'depress', 'anxiety', 'help', and 'can't', indicating a pervasive theme of psychological distress and the need for support. Topic 4 includes terms like 'love', 'told', 'lie', 'night', and 'make', which are consistent with romantic relationships marked by uncertainty or trust issues. Finally, Topic 5 concentrates on family- and marriage-related terms such as 'wife', 'family', 'love', and 'tell', supporting an interpretation centered on emotional communication within close and adult relationships. To further improve interpretability, we also include an alternative visualization based on unique top words for each topic in Section S6.1 of the SM. In this representation, shared high-frequency terms are assigned only to the topic where they have the largest probability, thereby highlighting topic-specific content more clearly.

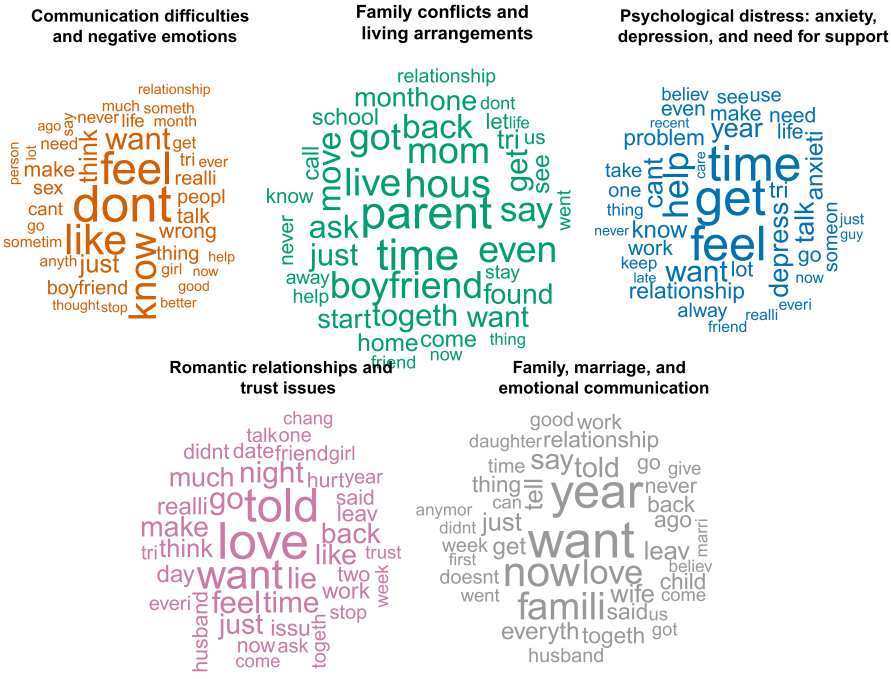

Fig. 6 Mental Health application: Word clouds representing the estimated word proportions $\hat{\phi}_k$ for each topic. Each cloud displays the 40 words with the largest estimated probability within that topic

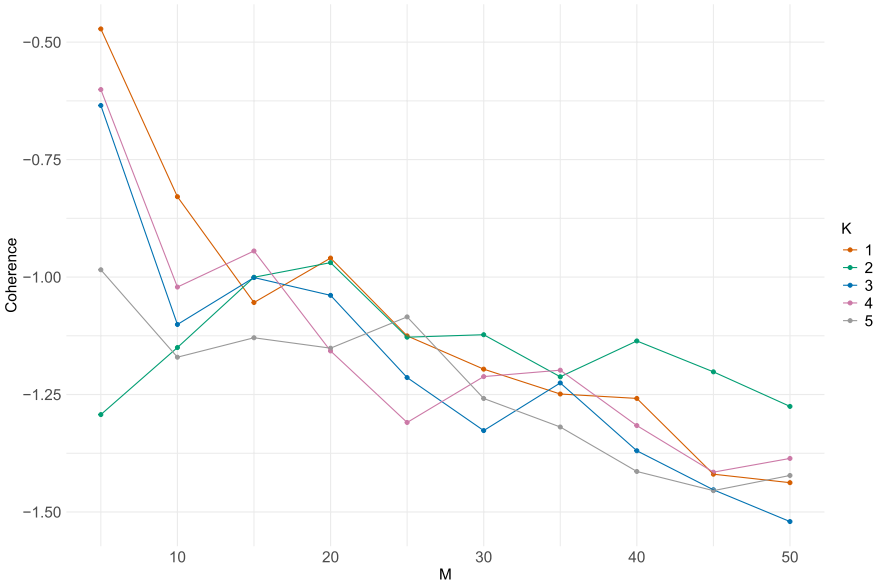

Fig. 7 Mental Health application: Coherence values for each detected topic as a function of M, the number of most probable words considered

Table 1 Mental Health application: Frequency of each detected sub-topic

Sub-topic	1-4	1-5	2-1	2-3	2-4	2-5	3-1	3-5	4-1	4-5	5-1	5-2
Freq	1	41	7	3	1	44	3	21	4	14	2	2

As shown in Fig. 6, Topic 2 (the green topic) is characterized by top words with similar sizes, and therefore with similar estimated probabilities. Topics exhibiting this pattern are typically regarded as poorly coherent, whereas a coherent topic is usually characterized by a small set of highly informative words. To quantify this aspect for each detected topic, we computed the coherence measure proposed by Mimno et al. (2011). Figure 7 reports the coherence values obtained by considering the M most probable words for each topic. For small values of M, Topic 2 clearly displays low coherence, confirming the insights suggested by the word clouds.

To deepen the analysis of the EFLDA model, we identified the sub-topics associated with each of the 143 questions as described in Sect. 4. Table 1 reports the distribution of the detected sub-topics. The questions covered 12 out of the $5!/(5-2)! = 20$ possible sub-topic labels, with the most frequent being "2–5" (family and living arrangements that may lead to anxiety and depression), "1–5" (communication difficulties within family and marriage), and "3–5" (psychological distress within one's marriage and family unit). It is interesting to note that we also observed sub-topics that differ in the importance of the same two concepts (i.e., sub-topics "1-4" and "4-1", "1-5" and "5-1", and "2-5" and "5-2").

Lastly, to investigate the ability of the EFLDA model to accommodate positively correlated topics, we fitted it using an alternative set of hyperparameters. Specifi-

Table 2 Mental Health application: Prior correlation matrix imposed on the elements of each $\boldsymbol{\theta}_d$ by considering a scenario in which one can expect positive correlations

	$\theta_{d,1}$	$\theta_{d,2}$	$\theta_{d,3}$	$\theta_{d,4}$	$\theta_{d,5}$
$\theta_{d,1}$	1	0.455	0.224	−0.149	−0.468
$\theta_{d,2}$	0.455	1	0.017	−0.216	−0.392
$\theta_{d,3}$	0.224	0.017	1	−0.294	−0.426
$\theta_{d,4}$	−0.149	−0.216	−0.294	1	−0.529
$\theta_{d,5}$	−0.468	−0.392	−0.426	−0.529	1

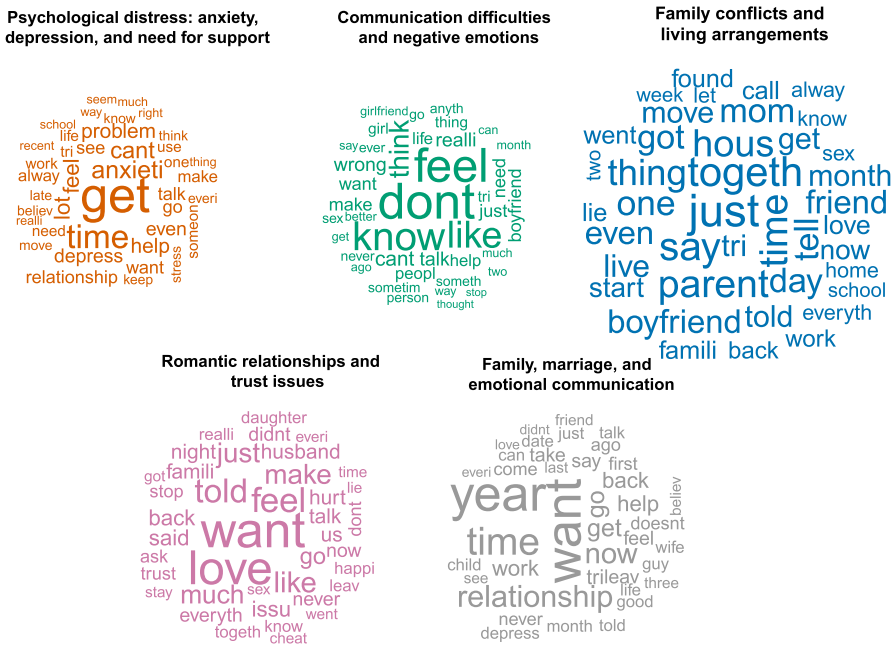

Fig. 8 Mental Health application (EFLDA with some positive prior correlations): Word clouds representing the estimated word proportions $\hat{\phi}_k$ for each topic. Each cloud displays the 40 words with the largest estimated probability within that topic

cally, assuming the availability of prior information or the need to impose a particular correlation structure on the latent topics, we retained the same α^+ and $\mathbf{p}$ suggested in Sect. 5, obtaining the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ by mapping back the choices $\boldsymbol{\mu} = (1/K, \ldots, 1/K)^\intercal$ and $\tilde{\mathbf{w}} = (0.05, 0.275, 0.423, 0.685, 0.95)^\intercal$. This configuration allowed us to have the correlation matrix shown in Table 2.

Interestingly, this choice did not alter the overall topic structure, as the word clouds in Fig. 8 closely resemble those in Fig. 6, the main difference being a permutation of the topics. Nonetheless, by modifying the hyperparameters, the topics exhibit greater coherence for small values of M (this is true for all five topics), as shown by Fig. 9.

Overall, the EFLDA model produces coherent and stable topic representations. The inferred topics are clearly interpretable and aligned with distinct thematic dimensions (e.g., psychological distress, communication difficulties, family conflicts, and relationship issues), while still allowing for controlled overlap when appropriate. This reflects the advantage of the probabilistic framework, which enforces a struc-

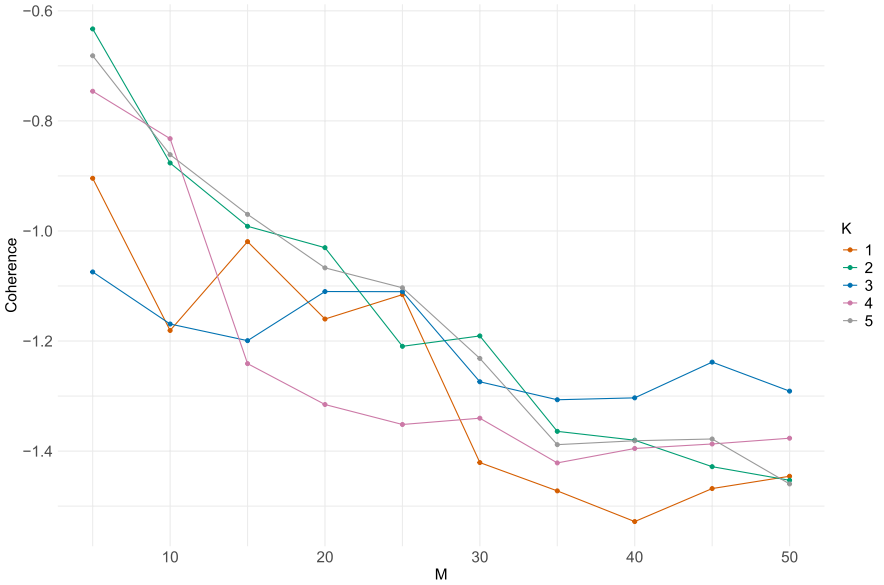

Fig. 9 Mental Health application (EFLDA with some positive prior correlations): Coherence values for each detected topic as a function of M, the number of most probable words considered

tured representation of topics and leads to more robust and reproducible results across different preprocessing settings.

6.2 Microbiome data

In the context of microbiome data analysis, latent topic models can be adopted to detect microbial communities, namely latent profiles of microbial abundance (Sankaran and Holmes 2019). Communities are defined as a unique composition of taxa that are expected to co-occur due to similar environmental and/or biological functions (Giampino et al. 2025). Within this domain, taxa are handled as words forming a vocabulary, and different sample sites (or sample units) are considered as the documents.

The COMBO dataset, originally introduced by (Wu et al. 2011), comes from a cross-sectional study of 98 healthy participants, where DNA extracted from stool samples was sequenced using 454/Roche technology targeting 16 S rRNA gene regions. The taxonomic profiling procedure identified 3, 068 operational taxonomic units (OTUs), which were then summarized into 87 genera observed in at least one sample. In this work, we restrict attention to $D = 96$ individuals, excluding two participants whose microbial profiles were incomplete. Sequencing depth showed notable variability across samples, with read counts ranging from 1, 242 to 14, 544. On average, each sample contained 6, 938 reads, with a standard deviation of 3, 024. For the purposes of this analysis, we concentrate on the $V = 45$ most prominent bacterial taxa in the dataset. Altogether, the dataset includes 666, 009 sequencing reads, with *Bacteroides* taxon emerging as the most prevalent genus, contributing 355, 943 reads.

Due to the relatively small value of D, splitting the corpus into separate training and test sets is not feasible. To identify the optimal number of communities K for LDA, FLDA, EFLDA, and NMF models, we therefore follow Giampino et al. (2025) and adopt a 4-fold cross-validation (CV) procedure. Figure 10 shows the median perplexity of the four models as a function of K.

The results suggest that, although EFLDA achieves the best performance on the training set, it only slightly outperforms FLDA at their shared optimal value $K = 5$. In contrast, LDA attains its lowest perplexity at $K = 7$, suggesting the need for a more complex model. The perplexity values for NMF on the test set are omitted for graphical reasons, as the method produces unstable (diverging) perplexity due to overfitting. To implement BERTopic and Top2Vec, we reconstructed the individual samples from the DTM, thereby preserving the original sequential ordering of the taxa. Furthermore, we evaluated a second configuration in which the taxa were randomly shuffled within each biological sample (i.e., we applied a random permutation to the taxa appearing multiple times). The corresponding results are detailed in Tables S10 and S11 of the SM. Under both settings, BERTopic consistently identifies only two communities. While this may suggest a parsimonious structure, a closer inspection of the top 10 taxa within each community (Table S10) reveals that the two communities are nearly indistinguishable, sharing nine of their ten most important taxa with only minor differences in ranking (e.g., swaps in some positions). This indicates a limited ability to capture distinct community structures. Moreover, the results are highly sensitive to the input representation, as shuffling the taxa order leads to noticeable changes in the identified communities even if the taxa (i.e., the words) are the same (Table S11). This behavior is likely due to both the nature of the data and the characteristics of the model: the corpus consists solely of taxa names, which lack the rich semantic structure typically exploited by embedding-based methods. As a result, BERTopic, which relies on preserving semantic relationships in the embedding space, struggles to identify meaningful and well-separated communities

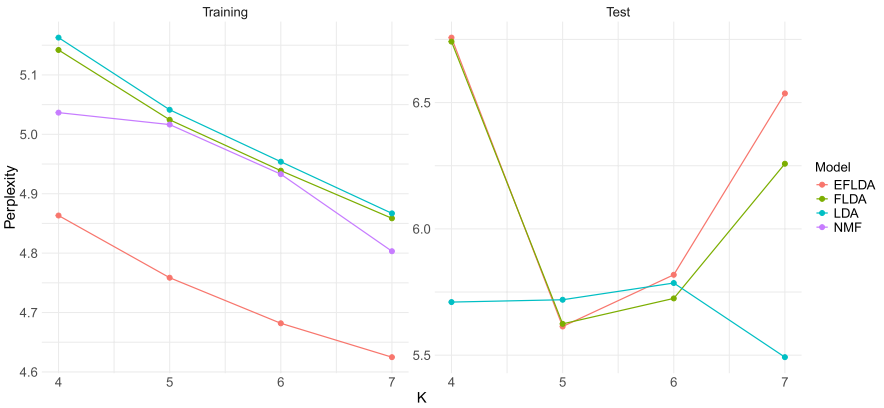

Fig. 10 COMBO application: Median of the perplexity value for the 4-folds CV on the training (left panel) and the test set (right panel) as a function of K. In the right panel, the NMF has been removed for graphical reasons

in this setting. Similarly, Top2Vec identifies only a single community, grouping all taxa together and failing to extract any meaningful structure.

In contrast, the EFLDA model is specifically designed to handle such data through its probabilistic framework, which does not rely on semantic embeddings but instead captures co-occurrence patterns and dependence structures. This allows EFLDA to identify well-separated and interpretable communities, demonstrating greater robustness and effectiveness in non-textual compositional settings such as microbiome data. Figure 11 shows the taxa clouds of the five selected communities detected by the EFLDA model once re-fitted on the whole corpus, which are broadly consistent with those obtained by Giampino et al. (2025), albeit with some differences. Table S12 reports the top 5 unique taxa for each community identified by the model.

Finally, we applied the sub-topic detection procedure described in Sect. 4. In the microbiome setting, sub-topics can be interpreted as finer structures within communities, reflecting variations of the same community driven by different factors, such as dietary regimes or environmental conditions. Table S13 in the SM reports the frequency distribution of the detected sub-topics. The three most frequent sub-topics are all characterized by the first community (the red one) as the dominant component (i.e., "1-2", "1-3", and "1-4"), while the fourth most frequent still involves that community, but as the second most prominent component (i.e., "2-1"). The remaining 12 detected sub-topics account for 34.4% of the biological samples.

6.3 BBC sports news

One of the most common applications of latent topic models concerns news data. In this application, we consider a collection of sports news articles from the BBC network, referring to the 2004/2005 season and covering five different sports: athletics, cricket, football, rugby, and tennis. Data consist in a DTM with 736 documents and a vocabulary of 540 unique words, obtained after removing words appearing fewer

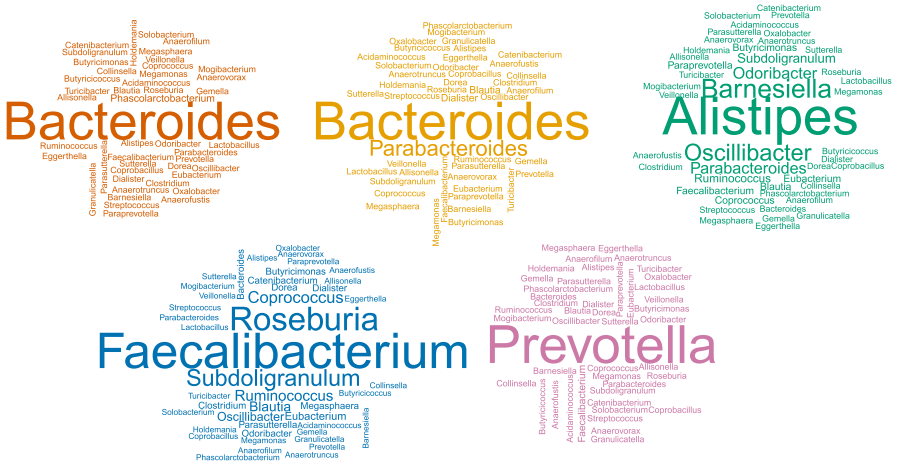

Fig. 11 COMBO application: Taxa clouds of the estimated proportions of taxa $\hat{\phi}_k$ for each latent community identified by EFLDA model

than 50 times. The total number of words is 72, 350, with an average of 98 words per article. Notably, by assuming that each topic should interpret one of the considered sports, the ground-truth number of topics in the original dataset is 5.

Figure 12 reports the perplexity values computed on the training and test sets obtained via an 80%-20% random split of the original dataset. Results for NMF on the test set are omitted, as the method exhibits severe overfitting on the training data and produces diverging perplexity values on the test set, with magnitudes up to 10^3 higher than those of the other models. Among the probabilistic models, EFLDA achieves the lowest perplexity, with an optimal number of $K = 6$ topics. In contrast, BERTopic identifies a substantially larger number of topics, detecting 11 topics when applied to documents reconstructed from the DTM (thus preserving word order), and 16 topics when applied to documents with shuffled word order as done in Sect. 6.2. In both cases, the number of identified topics is twice or thrice the ground-truth number of topics. Top2Vec, instead, shows results more consistent with the probabilistic models, identifying 4 or 5 topics depending on whether the documents preserve the original word order or are shuffled (Tables S14 and S15 in the SM). This result is not surprising, since Top2Vec is known to perform better in settings with larger corpora and richer semantic information.

The six latent topics identified by the EFLDA model, once re-fitted on the whole corpus (Table 3), reveal a set of thematically interpretable but partially overlapping semantic clusters. Topic 2 is strongly characterized by performance-oriented terminology (e.g., '100 m', 'hurdler', 'sprint', and 'run'), suggesting a focus on competitive events and measurable athletic performance. Topic 3 complements this by incorporating elements related to preparation, qualification, and international context (e.g., 'preparation', 'qualification', 'Nation', 'Germany', and 'Sweden'), indicating a broader narrative around competitive participation and event organization. Topic 5 emphasizes achievement and competitive progression, with terms such as 'reach', 'determination', 'outdoor', and 'title', reflecting success and advancement in structured events. In contrast, Topic 1 captures more affective and experiential dimensions, including terms like 'frustration', 'effort', 'quit', and "impress," which point

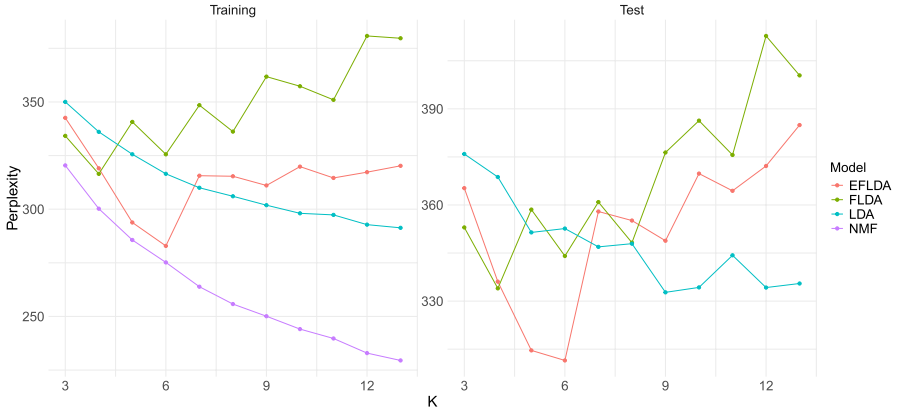

Fig. 12 BBC Sport application: Perplexity value on the training (left panel) and the test set (right panel) as a function of K. In the right panel, the NMF has been removed for graphical reasons

Table 3 BBC Sport application: Top 5 unique representative words for each of the six latent topics identified by the EFLDA model

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
shook	big	Sweden	past	person	March
heat	100 m	18	form	sprint	just
frustration	hurdler	preparation	came	compro-mise	airport
aim	Shevchen-ko	venue	chance	Tuesday	go
line	hunt	week	record	reach	boost
quit	Sonia	cragg	season	AAA	before
break	run	pace	60 m	Scott	star
impress	come	qualification	confid	determi-nation	titl
Irina	Nation	seven	25year-old	place	under-stand
effort	Germany	sai	runner	outdoor	Phillip

to personal narratives and emotional responses. Topic 4 appears to aggregate general descriptors of performance and temporal progression (e.g., 'past', 'record', 'season', and 'confidence'), suggesting a focus on evaluation and longitudinal performance across several disciplines. Finally, Topic 6 is dominated by temporal and contextual markers such as 'march', 'Tuesday', 'airport', and 'before', which are indicative of event timing, logistics, and reporting context rather than domain-specific content. It is noteworthy that the identified topics, while numerically consistent with the number of "ground truth" categories (i.e., the five sports), reveal a latent structure that diverges from the original labels. This finding underscores the utility of latent topic models in high-dimensional settings, as they can uncover underlying co-occurrence patterns that might otherwise remain imperceptible through traditional categorical analysis. In Figure S16 of the SM, we report the coherence values obtained by considering the M most probable words for each topic.

7 Conclusions

In this work, we introduced the EFLDA, a probabilistic topic model designed to address key limitations of classical LDA in capturing complex dependence structures within textual data. Unlike LDA, which relies on a Dirichlet prior for the vector of topic proportions, EFLDA employs a more flexible prior that allows both positive and negative correlations on the simplex. Notably, LDA and FLDA emerge as special cases of EFLDA, highlighting the generality of the proposed framework. Through its finite mixture structure and additional hyperparameters, the model also enables model-based clustering and the detection of sub-topics, an aspect particularly relevant in domains where thematic distinctions are subtle and highly context-dependent.

We developed a CGS algorithm that exploits the conjugacy properties of the EFD distribution with respect to multinomial and categorical likelihoods, providing tractable and efficient posterior inference. We also carried out simulation studies to compare three initialization methods for the model's hyperparameters.

The application across a range of contexts highlights the superiority of EFLDA over competing models, both in terms of model fit and in its ability to uncover meaningful topic and subtopic structures. In particular, compared with recent non-probabilistic embedding-based approaches such as BERTopic and Top2Vec, EFLDA yields more stable and interpretable topic representations. While embedding-based models are effective in capturing semantic relationships, they are often sensitive to preprocessing choices and input structure, which may lead to overlapping topics, unstable clusters, or the presence of outlier components (Egger and Yu 2022). Such issues are further exacerbated when the raw data lack an inherent semantic structure. In contrast, the probabilistic framework underlying EFLDA enables consistent estimation of topic-word and topic-document distributions, as well as principled uncertainty quantification by supporting (Bayesian) inference. These findings highlight the model's potential in computational social science and digital mental health research, where the identification of latent topics plays a central role in understanding human behavior and informing the design of targeted interventions, as well as in microbiome studies, where observations do not possess an intrinsic semantic structure.

Future work will explore further extensions of the EFD prior by relaxing assumptions on the parameters ϕ_k governing topic compositions. This would enable the model to capture more complex relationships among the elements of each latent structure.

Additional research directions include hierarchical or nested formulations to capture dependencies among sub-topics, as well as temporal extensions to model topic evolution.

Together, these extensions would further increase the flexibility and applicability of EFLDA, enabling the discovery of richer latent structures across diverse scientific domains.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-026-00690-9.

Acknowledgements We greatly acknowledge the DEMS Data Science Lab for supporting this work by providing computational resources.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. This research was financially supported by the University of Milano-Bicocca.

Data Availability The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest All authors declare no conflict of interest for this article.

Informed consent Not applicable to this study.

References

Ahmed F, Nawaz M, Jadoon A (2022) Topic Modeling of the Pakistani Economy in English Newspapers via Latent Dirichlet Allocation (LDA). SAGE Open 12(1):1–14. https://doi.org/10.1177/21582440221079931

Aitchison J (2003) The Statistical Analysis of Compositional Data, 2nd edn. The Blackburn Press, London

Angelov D (2020) Top2Vec: distributed Representations of Topics. arXiv preprint arXiv:2008.09470

Arseniev-Koehler A, Cochran V, Mays SD et al (2022) Integrating topic modeling and word embedding to characterize violent deaths. Proceedings of the National Academy of Sciences (PNAS) 119(10):e2108801119. https://doi.org/10.1073/pnas.2108801119

Ascari R, Di Brisco AM, Migliorati S et al (2024) A multivariate mixture regression model for constrained responses. Bayesian Anal 19(2):377–405. https://doi.org/10.1214/22-BA1359

Bittermann A, Fischer A (2018) How to identify hot topics in psychology using topic modeling. Zeitschrift für Psychologie 226(1):3–13. https://doi.org/10.1027/2151-2604/a000318

Blei DM, Lafferty JD (2007) A correlated topic model of Science. The Annals of Applied Statistics 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

Breuninger TA, Wawro N, Breuninger J, et al (2021) Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. Microbiome 9(61). https://doi.org/10.1186/s40168-020-00969-9

Carron-Arthur B, Reynolds J, Bennett K, et al (2016) What's all the talk about? Topic modelling in a mental health Internet support group. BMC Psychiatry 16(367). https://doi.org/10.1186/s12888-016-1073-5

Couto M, Perez A, Parapar J et al (2025) Temporal word embeddings for early detection of psychological disorders on social media. Journal of Healthcare Informatics Research. https://doi.org/10.1007/s41666-025-00186-9

Dao B, Nguyen T, Venkatesh S et al (2017) Latent sentiment topic modelling and nonparametric discovery of online mental health-related communities. International Journal of Data Science and Analytics 4:209–231. https://doi.org/10.1007/s41060-017-0073-y

Deek R, Li H (2021) A zero-inflated latent Dirichlet allocation model for microbiome studies. Frontiers Genetics 11:602594. https://doi.org/10.3389/fgene.2020.602594

Egger R, Yu J (2022) A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. Front Sociol 7:886498. https://doi.org/10.3389/fsoc.2022.886498

Feuerriegel S, Ratku A, Neumann D (2016) Analysis of how underlying topics in financial news affect stock prices using latent Dirichlet allocation. In: 2016 49th Hawaii International Conference on System Sciences (HICSS), pp 1072–1081, https://doi.org/10.1109/HICSS.2016.137

Finch WH, Hernández Finch ME, McIntosh CE et al (2018) The use of topic modeling with latent Dirichlet analysis with open-ended survey items. Translational Issues in Psychological Science 4(4):403–424. https://doi.org/10.1037/tps0000173

Gao X, Sazara C (2023) Discovering mental health research topics with topic modeling.arxiv:2308.13569

Giampino A, Ascari R, Migliorati S (2025) A flexible mixed-membership model for community and enterotype detection for microbiome data. Computational Statistics & Data Analysis 210:108181. https://doi.org/10.1016/j.csda.2025.108181

Greene D, O'Callaghan D, Cunningham P (2014) How many topics? Stability analysis for topic models. In: Joint European conference on machine learning and knowledge discovery in databases, Springer, pp 498–513, https://doi.org/10.1007/978-3-662-44848-9_32

Griciūtė B, Han L, Nenadic G (2023) Topic modelling of Swedish newspaper articles about Coronavirus: a case study using latent Dirichlet allocation method. In: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), pp 627–636,https://doi.org/10.1109/ICHI57859.2023.00110

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl. 1):5228–5235. https://doi.org/10.1073/pnas.0307752101

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure arXiv:2203.05794 arXiv preprint

Hagg LJ, Merkouris SS, O'Dea GA et al (2022) Examining analytic practices in latent Dirichlet allocation within psychological science: Scoping review. J Med Internet Res 24(11):e33166. https://doi.org/10.2196/33166

Higashi K, Suzuki S, Kurosawa S et al (2018) Latent environment allocation of microbial community data. PLoS Comput Biol 14(6):e1006143. https://doi.org/10.1371/journal.pcbi.1006143

Hosoda S, Nishijima S, Fukunaga T et al (2020) Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation. Microbiome 8:95. https://doi.org/10.1186/s40168-020-00864-3

Jelodar H, Wang Y, Yuan C et al (2019) Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia tools and applications 78:15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Jones N, Jaques N, Pataranutaporn P, et al (2019) Analysis of online suicide risk with document embeddings and latent Dirichlet allocation. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp 1–5, https://doi.org/10.1109/ACIIW.2019.8925077

Kuang D, Choo J, Park H (2014) Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Partitional clustering algorithms. Springer, p 215–243, https://doi.org/10.1007/978-3-319-09259-1_7

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86. https://doi.org/10.1214/aoms/1177729694

LeBlanc P, Ma L (2023) Microbiome subcommunity learning with logistic-tree normal latent Dirichlet allocation. Biometrics 79(3):2321–2332. https://doi.org/10.1111/biom.13772

Liu JS (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J Am Stat Assoc 89(427):958–966. https://doi.org/10.1080/01621459.1994.10476829

Liu S, Zhang RY, Kishimoto T (2021) Analysis and prospect of clinical psychology based on topic models: hot research topic and scientific trends in the latest decades. Psychiology Health & Medicine 26(4):395–407. https://doi.org/10.1080/13548506.2020.1738019

Mekaoui S, Chaker I, Zarghili A et al (2025) Systematic literature review of topic labeling. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3573521

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, UK., pp 262–272, https://aclanthology.org/D11-1024/

Ongaro A, Migliorati S (2013) A generalization of the Dirichlet distribution. J Multivar Anal 114(1):412–426. https://doi.org/10.1016/j.jmva.2012.07.007

Ongaro A, Migliorati S, Ascari R (2020) A new mixture model on the simplex. Stat Comput 30:749–770. https://doi.org/10.1007/s11222-019-09920-x

Paul MJ, Dredze M (2014) Discovering health topics in social media using topic models. PLoS One 9(8). https://doi.org/10.1371/journal.pone.0103408

Salmi S, van der Mei R, Mérelle S et al (2024) Topic modeling for conversations for mental health helplines with utterance embedding. Telematics and Informatics Reports 13:100126. https://doi.org/10.1016/j.teler.2024.100126

Sankaran K, Holmes SP (2019) Latent variable modeling for the microbiome. Biostatistics 20(4):599–614. https://doi.org/10.1093/biostatistics/kxy018

Schiavon L (2025) Addressing topic modelling via reduced latent space clustering. Statistical Methods & Applications 34:1–20. https://doi.org/10.1016/j.teler.2024.100126

Stan Development Team (2022) Stan modeling language users guide and reference manual, version 2.32. https://mc-stan.org

Subeno B, Kusumaningrum R, et al (2018) Optimisation towards latent Dirichlet allocation: its topic number and collapsed Gibbs sampling inference process. International Journal of Electrical & Computer Engineering (2088-8708) 8(5). https://doi.org/10.11591/ijece.v8i5.pp3204-3213

Wallach H, Mimno D, McCallum A (2009) Rethinking LDA: Why priors matter. Adv Neural Inf Process Syst 22. https://doi.org/10.5555/2984093.2984314

Westrupp EM, Greenwood CJ, Fuller-Tyszkiewicz M et al (2022) Text mining of Reddit posts: using latent Dirichlet allocation to identify common parenting issues. PLoS ONE 17(2):e0262529. https://doi.org/10.1371/journal.pone.0262529

Wu GD, Chen J, Hoffmann C et al (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334(6052):105–108. https://doi.org/10.1126/science.1208344

Xin Y, Grabowska ME, Gangireddy S et al (2025) Improving topic modeling performance on social media through semantic relationships within biomedical terminology. PLoS ONE 20(2):e0318702. https://doi.org/10.1371/journal.pone.0318702

Xue J, Chen J, Chen C et al (2020) Public discourse and sentiment during the COVID 19 pandemic: Using latent Dirichlet allocation for topic modeling on Twitter. PLoS ONE 15(9):e0239441. https://doi.org/10.1371/journal.pone.0239441

Yao L, Mimno D, McCallum A (2009) Efficient methods for topic model inference on streaming document collections. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 937–946, https://doi.org/10.1145/1557019.1557121

Yin R, Tian R, Wu J et al (2022) Exploring the factors associated with mental health attitude in China: A structured topic modeling apporach. Int J Environ Res Public Health 19(19):12579. https://doi.org/10.3390/ijerph191912579