Clustering Count Data with Stochastic Expectation Propagation

Xavier Sumba[1](✉), Nuha Zamzami[2,3], and Nizar Bouguila[3]

[1] Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada
xavier.sumba93@ucuenca.edu.ec

[2] Department of Computer Science and Artificial Intelligent, University of Jeddah, Jeddah, Saudi Arabia

[3] Concordia Institute for Information Systemes Engineering, Concordia University, Montreal, Canada

Abstract. Clustering count vectors is a challenging task given their sparsity and high-dimensionality. An efficient generative model called EMSD has been recently proposed, as an exponential-family approximation to the Multinomial Scaled Dirichlet distribution, and has shown to offer excellent modeling capabilities in the case of sparse count data and to overcome some limitations of the frameworks based on the Dirichlet distribution. In this work, we develop an approximate Bayesian learning framework for the parameters of a finite mixture of EMSD using the Stochastic Expectation Propagation approach. In this approach, we maintain a global posterior approximation that is being updated in a local way, which reduces the memory consumption, important when making inference in large datasets. Experiments on both synthetic and real count data have been conducted to validate the effectiveness of the proposed algorithm in comparison to other traditional learning approaches. Results show that SEP produces comparable estimates with traditional approaches.

Keywords: Mixture model · Emsd distribution · Stochastic expectation propagation

1 Introduction

Statistical methods are excellent at modeling semantic content of text documents [9]. More specifically, document clustering is widely used in a variety of applications such as text retrieval or topic modeling, (see e.g. [3]). Words in text documents usually exhibit appearance dependencies, *i.e.*, if word w appears once, it is more probable that the same word w will appear again. This phenomenon is called burstiness, which has shown to be addressed by introducing the prior information into the construction of the statistical model to obtain several computational advantages [15]. Given that the Dirichlet distribution is

N. T. Nguyen et al. (Eds.): ACIIDS 2021, LNAI 12672, pp. 119–129, 2021.
https://doi.org/10.1007/978-3-030-73280-6_10

generally taken as a conjugate prior to the multinomial, the most popular hierarchical approach is the Dirichlet Compound Multinomial (DCM) distribution [14]. While the Multinomial distribution fails to model the words burstiness given its dependency assumption, the DCM distribution not only captures this behavior but also models text data better [14]. However, The Dirichlet distribution has its own limitations due to is negative covariance structure and equal confidence [11,24]. Hence, a generalization of it called the Scaled Dirichlet (SD) distribution has shown to be a good alternative as a prior to the multinomial resulting in the Multinomial scaled Dirichlet (MSD) distribution recently proposed in [25]. Indeed, MSD has shown to have high flexibility in count data modeling with superior performance in several real-life challenging application [25–28]. Despite its flexibility, MSD distribution shares similar limitations to the one with DCM since its parameter estimation is slow, especially in high-dimensional spaces. Thus, [28] proposed a close exponential-family approximation called EMSD to combine the flexibility and efficiency of MSD with the desirable statistical and computational properties of the exponential family of distributions, including sufficiency. EMSD has shown to reduce the complexity and computational efforts, considering the sparsity and high-dimensionality nature of count data.

In this work, we study the application of the Bayesian framework for learning the exponential-family approximation to the Multinomial Scaled Dirichlet (EMSD) mixture model which has been shown to be an appropriate distribution to model the burstiness in high-dimensional feature space. In particular, we propose a learning approach for an EMSD mixture model using Stochastic Expectation Propagation (SEP) [10] for parameter estimation. Indeed, SEP combines both Assumed Density Filtering (ADF) and Expectation Propagation (EP) in order to scale to large datasets while mantaining accurate estimations. Only EP is usually more accurate than methods such as variational inference and MCMC [1,18], and SEP solves some of the problems encountered when using EP given that the number of parameters increase according to number of datapoints. Thus, SEP is a deterministic approximate inference method that prevents memory overheads when increasing the number of data points. EP has shown to be an appropriate generalization in the case of Gaussian mixture model [20], hierarchical models such as LDA [18] or even infinite mixture models [6]. Furthermore, SEP has been used with Deep Gaussian process [4], showing the benefits of scalable Bayesian inference and outperforming traditional Gaussian process. The contributions of this paper are summarized as follows: 1) we show that SEP can provide effective parameter estimates when dealing with large datasets; 2) we derive foundations to learn an EMSD mixture model using SEP; 3) we exhaustively evaluate the proposed approach on synthetic and real count data and compare the performance with other models and learning approaches.

2 The Exponential-Family Approximation to MSD Distribution

In the clustering setting, We are given a dataset $\mathcal{X}$ with D samples $\mathcal{X} = \{\mathbf{x_i}\}_{i=1}^{D}$, each $\mathbf{x}_i$ is a vector of count data (*e.g.* a text document or an image, represented as a frequencies vector of words or visual words, respectively). We assume that each data set has a vocabulary of size V.

The the Multinomial Scaled Dirichlet (MSD) is the marginal distribution defined by integrating out the probability parameter of scaled Dirichlet over all possible multinomials, and it is given by [25]:

$$\mathcal{MSD}(\mathbf{x} \mid \boldsymbol{\rho}, \boldsymbol{\nu}) = \frac{n!}{\prod_{w=1}^{V} x_w!} \frac{\Gamma(s)}{\Gamma(s+n) \prod_{w=1}^{V} \nu_w^{x_w}} \prod_{w=1}^{V} \frac{\Gamma(x_w + \rho_w)}{\Gamma(\rho_w)} \tag{1}$$

Note that the authors in [25] use the approximation $\left(\sum_{w=1}^{V} \nu_w p_w\right)^{\sum_{w=1}^{V} x_w} \approx \prod_{w=1}^{V} \nu_w^{x_w}$. It is worth mentioning that DCM is a special case of MSD, such that when $\boldsymbol{\nu} = 1$ in Eq. (1), we obtain the Dirichlet Compound Multinomial (DCM) distribution [14]. Similar to DCM, the considered model MSD, has an intuitive interpretation representing the Scaled Dirichlet as a general topic and the Multinomial as a document-specific subtopic, making some words more likely in a document $\mathbf{x}$ based on word counts.

The representation of text documents is very sparse as many words in the vocabulary do not appear in most of the documents. Thus, in [28], the authors note that using only the non-zero values of $\mathbf{x}$ is computationally efficient since $x_w! = 1$, $\nu_w^{x_w} = 1$ and $\Gamma(x_w + \rho_w)/\Gamma(\rho_w) = 1$ when $x_w = 0$. Moreover, since in high dimensional data the parameters are very small, [5], the following fact for small values of ρ when $x \geq 1$ was used in [28]:

$$\lim_{\rho \to 0} \frac{\Gamma(x+\rho)}{\Gamma(\rho)} - \Gamma(x)\rho = 0 \tag{2}$$

Thus, being able to approximate $\Gamma(x_w + \rho_w)/\Gamma(\rho_w) = \Gamma(x_w)\rho_w$ and using the fact that $\Gamma(x_w) = (x_w - 1)!$ leads to an approximation of the MSD distribution known as the Exponential-family approximation to the MSD distribution (EMSD), given by:

$$\mathcal{EMSD}(\mathbf{x} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{n!}{\prod_{w:x_w \geq 1}^{V} x_w} \frac{\Gamma(s)}{\Gamma(s+n)} \prod_{w:x_w \geq 1}^{V} \frac{\alpha_w}{\beta_w^{x_w}} \tag{3}$$

The parameters of the EMSD distribution are denoted by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to distinguish them from the MSD parameters for clarity.

3 Stochastic Expectation Propagation

Efficient inference and learning for probabilistic models that scale to large datasets are essential in the Bayesian setting. Thus, a variety of methods have been proposed from sampling approximations [17] to distributional approximations such as stochastic variational inference [8].

Another deterministic approach is Expectation Propagation (EP) that commonly provides more accurate approximations compared to sampling methods [21] and variational inference [19,20]. Yet, the number of parameters grows with the number of data points, causing memory overheads and making it difficult to scale to large datasets. Besides, Assumed Density Filtering (ADF) [22], which has been introduced before EP, maintains a global approximating posterior; however, it results in poor estimates. Therefore, [10] proposed an alternative to push EP to large datasets denominated Stochastic Expectation Propagation (SEP). SEP takes the best of these two methods by maintaining a global approximation that is updated locally. It does this by introducing a global site that captures the average effect of the likelihood sites and, as a result avoiding memory overheads.

Given a probabilistic model $p(\mathcal{X} \mid \boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$ drawn from a prior $p_0(\boldsymbol{\theta})$, SEP approximates a target distribution $p(\boldsymbol{\theta} \mid \mathcal{X})$, which is commonly the posterior, with a global approximation $q(\boldsymbol{\theta})$ that belongs to the exponential family. The target distribution must be factorizable such that the posterior can be split into D sites $p(\boldsymbol{\theta} \mid \mathcal{X}) \propto p_0(\boldsymbol{\theta}) \prod_{i=1}^{D} p_i(\boldsymbol{\theta})$; the initial site p_0 is commonly interpreted as the prior distribution and the remaining p_i sites represent the contribution of each ith item to the likelihood. The approximating distribution must admit a similar factorization, $q(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta})\tilde{p}(\boldsymbol{\theta})^D$.

Unlike EP, the SEP maintains a global approximating site, $\tilde{p}(\boldsymbol{\theta})^D$, to capture the average effect of a likelihood on the posterior. Thus, we only have to maintain the parameters of the approximate posterior and approximate global site that commonly belongs to the exponential family. Consequently, each site is refined to create a cavity distribution by dividing the global approximation over one of the copies of the approximate site, $q^{\setminus 1}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})/\tilde{p}(\boldsymbol{\theta})$.

Additionally, in order to approximate each site, a new tilted distribution is introduced using the cavity distribution and the current site $\hat{p}_i(\boldsymbol{\theta}) \propto p_i(\boldsymbol{\theta})q^{\setminus 1}(\boldsymbol{\theta})$.

Subsequently, a new posterior is found by minimizing the Kullback Leibler divergence $D_{KL}(\hat{p}_i(\boldsymbol{\theta}) \mid\mid q^{new}(\boldsymbol{\theta}))$ such that $\tilde{p}_i(\boldsymbol{\theta}) \approx p_i(\boldsymbol{\theta})$. This minimization is equivalent to match the moments of those distributions [1,20]. Finally, the revised approximate site is updated by removing the remaining terms from the current approximation by employing damping [7,18] in order to make a partial update since $\tilde{p}_i$ captures the effect of a single likelihood function:

$$\tilde{p}(\boldsymbol{\theta}) = \tilde{p}(\boldsymbol{\theta})^{1-\eta} \left(\frac{q^{new}(\boldsymbol{\theta})}{q^{\setminus w}(\boldsymbol{\theta})} \right)^{\eta} = \tilde{p}(\boldsymbol{\theta})^{1-\eta} \tilde{p}_i(\boldsymbol{\theta})^{\eta} \tag{4}$$

Notice that η is the step size, and when $\eta = 1$, no damping is applied. A natural choice is $\eta = 1/D$.

4 EMSD Mixture Model

4.1 Clustering Model

We assume that we are given D documents drawn from a finite number of EMSD distributions, and each $\mathbf{x}_i$ document is composed of V words. $K \geq 1$ represents the number of mixture components. Thus, a document is drawn from its respective component j as follows: $\mathbf{x_i} \sim \mathcal{EMSD}(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$.

In a mixture model, a latent variable $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^{D}$ is introduced for each $\mathbf{x}_i$ document in order to represent the component assignment. We posit a Multinomial distribution for the component assignment such that $\mathbf{z}_i \sim Mult(1, \boldsymbol{\pi})$ where $\boldsymbol{\pi} = \{\pi_j\}_{j=1}^{K}$ represents the mixing weights, and they are subject to the constraints $0 < \pi_j < 1$ and $\sum_j \pi_j = 1$. In other words, $\mathbf{z}_i$ is a K-dimensional indicator vector containing a value of one when document $\mathbf{x}_i$ belongs to the component j, and zero otherwise. Note that in this setting the value of $z_{ij} = 1$ acts as the selector of the component that generates $\mathbf{x}_i$ document with parameters $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$; hence, $p(\mathbf{z}_i \mid \boldsymbol{\pi}) = \pi_j$. Thus, the full posterior is in Eq. 5.

$$p(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathcal{X}) \propto p(\boldsymbol{\pi})p(\boldsymbol{\alpha})p(\boldsymbol{\beta}) \prod_{i}^{D} \sum_{j}^{K} \pi_j p(\mathbf{x}_i \mid \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) \tag{5}$$

4.2 Parameter Learning

We use SEP in order to learn the parameters of the mixture model. We start by partitioning the likelihood in D sites and define a global approximating site for each of the latent variables ($\boldsymbol{\pi}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$). Theoretically, any distribution belonging to the exponential family can be used for the sites. We use a Gaussian distribution for the parameters of the EMSD distribution in order to facilitate calculations [12]. For the mixture weights, we use a Dirichlet distribution since it belongs to the $K-1$ simplex and fits the constraints imposed by the mixing weights. Equations 6 illustrate the choices for the approximate sites.

$$\tilde{p}(\boldsymbol{\pi}) \propto \prod_{j} \pi_j^{a_j} \quad \tilde{p}(\boldsymbol{\alpha}) = \prod_{j}^{K} \mathcal{N}(\boldsymbol{\alpha}_j \mid \boldsymbol{m}_j, p_j^{-1}) \quad \tilde{p}(\boldsymbol{\beta}) = \prod_{j}^{K} \mathcal{N}(\boldsymbol{\beta}_j \mid \boldsymbol{n}_j, q_j^{-1}) \tag{6}$$

Once the global approximate site has been defined, we compute the approximate posterior $q(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by introducing the priors and the average effect of the global site:

$$q(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto p(\boldsymbol{\pi}, \boldsymbol{a}^0)\tilde{p}(\boldsymbol{\pi} \mid \boldsymbol{a})^D \prod_{j}^{K} p\left(\boldsymbol{\alpha}_j \mid \boldsymbol{m}_j^0, (p_j^0)^{-1}\right) \tilde{p}\left(\boldsymbol{\alpha}_j \mid \boldsymbol{m}_j, (p_j)^{-1}\right)^D$$
$$p\left(\boldsymbol{\beta}_j \mid \boldsymbol{n}_j^0, (q_j^0)^{-1}\right) \tilde{p}\left(\boldsymbol{\beta}_j \mid \boldsymbol{n}_j, q_j^{-1}\right)^D$$

The approximate posterior distribution has the following parameters:

$$\boldsymbol{a}' = 1 + \boldsymbol{a}^0 + D\boldsymbol{a} \quad (p_j^{'})^{-1} = (p_j^0 + Dp_j)^{-1} \quad (q_j^{'})^{-1} = (q_j^0 + Dq_j)^{-1}$$
$$\boldsymbol{m}_j^{'} = (p_j^{'})^{-1}(p_j^0 \boldsymbol{m}_j^0 + Dp_j \boldsymbol{m}_j) \quad \boldsymbol{n}_j^{'} = (q_j^{'})^{-1}(q_j^0 \boldsymbol{n}_j^0 + Dq_j \boldsymbol{n}_j) \tag{7}$$

Consequently, we introduce a cavity distribution by removing the contribution of one of the copies of the global site. The cavity distribution has parameters

$\boldsymbol{a}^{\setminus 1}$, $\left(p_j^{\setminus 1}\right)^{-1}$, $\boldsymbol{m}_j^{\setminus 1}$, $\left(q_j^{\setminus 1}\right)^{-1}$, and $\boldsymbol{n}_j^{\setminus 1}$ illustrated in Eq. 8 that are calculated as follows: $q(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})/\tilde{p}_i(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$

$$\begin{gathered}\boldsymbol{a}^{\setminus 1} = \boldsymbol{a}^{'} - \boldsymbol{a} \quad \left(p_j^{\setminus 1}\right)^{-1} = \left(p_j^{'} - p_j\right)^{-1} \quad \left(q_j^{\setminus 1}\right)^{-1} = \left(q_j^{'} - q_j\right)^{-1} \\ \boldsymbol{m}_j^{\setminus 1} = \left(p_j^{\setminus 1}\right)^{-1}\left(p_j^{'}\boldsymbol{m}_j^{'} - p_j\boldsymbol{m}_j\right) \quad \boldsymbol{n}_j^{\setminus 1} = \left(q_j^{\setminus 1}\right)^{-1}\left(q_j^{'}\boldsymbol{n}_j^{'} - q_j\boldsymbol{n}_j\right) \end{gathered} \tag{8}$$

We use the cavity distribution and incorporate the *ith* site, resulting in the tilted distribution $\hat{p} = \frac{1}{Z_i} p_i q^{\setminus 1}$. We use this distribution to compute the KL divergence with the approximate distribution, which is equivalent to matching the moments. However, in this case, matching the moments leads to another analytically intractable integral (i.e. $Z_i = \sum_j^K \frac{a_j^{\setminus 1}}{\sum_k^K a_k^{\setminus 1}} \mathbb{E}_{p(\alpha_j, \beta_j)}\left[p(x_i \mid \alpha_j, \beta_j)\right]$). Thus, we compute this integral via Monte Carlo sampling. After matching the moments, we obtain the parameters for an updated approximate posterior.

$$\begin{gathered}\Psi(a_j^{'}) - \Psi(\sum_j^K a_j^{'}) = \Psi(a_j^{\setminus 1}) - \Psi(\sum_j^K a_j^{\setminus 1}) + \nabla_{a_j^{\setminus 1}} \log Z_i \\ p_j^{'} = \left(p_j^{\setminus 1}\right)^{-1}\left(2\nabla_{\left(p_j^{\setminus 1}\right)^{-1}} \log Z_i + p_j^{\setminus 1}\right)\left(p_j^{\setminus 1}\right)^{-1} - \left(\boldsymbol{m}_j^{'} - \boldsymbol{m}_j^{\setminus 1}\right)\left(\boldsymbol{m}_j^{'} - \boldsymbol{m}_j^{\setminus 1}\right)^{\intercal} \\ q_j^{'} = \left(q_j^{\setminus 1}\right)^{-1}\left(2\nabla_{\left(q_j^{\setminus 1}\right)^{-1}} \log Z_i + q_j^{\setminus 1}\right)\left(q_j^{\setminus 1}\right)^{-1} - \left(\boldsymbol{n}_j^{'} - \boldsymbol{n}_j^{\setminus 1}\right)\left(\boldsymbol{n}_j^{'} - \boldsymbol{n}_j^{\setminus 1}\right)^{\intercal} \\ \boldsymbol{m}_j^{'} = \boldsymbol{m}_j^{\setminus 1} + \left(p_j^{\setminus 1}\right)^{-1} \nabla_{\boldsymbol{m}_j^{\setminus 1}} \log Z_i \quad \boldsymbol{n}_j^{'} = \boldsymbol{n}_j^{\setminus 1} + \left(q_j^{\setminus 1}\right)^{-1} \nabla_{\boldsymbol{n}_j^{\setminus 1}} \log Z_i \end{gathered} \tag{9}$$

The values of $\boldsymbol{a}^{'}$ are calculated using fixed point iteration as described in [16]. Using this updated approximate posterior, we remove the cavity distribution in order to obtain an approximation to the *ith* site.

$$\begin{gathered}\boldsymbol{a} = \boldsymbol{a'} - \boldsymbol{a}^{\setminus 1} \quad (p_j)^{-1} = (p_j^{'} - p_j^{\setminus 1})^{-1} \quad \boldsymbol{m}_j = (p_j)^{-1}\left(p_j^{'}\boldsymbol{m}_j^{'} - p_j^{\setminus 1}\boldsymbol{m}_j^{\setminus 1}\right) \\ (q_j)^{-1} = (q_j^{'} - q_j^{\setminus 1})^{-1} \quad \boldsymbol{n}_j = (q_j)^{-1}\left(q_j^{'}\boldsymbol{n}_j^{'} - q_j^{\setminus 1}\boldsymbol{n}_j^{\setminus 1}\right) \end{gathered} \tag{10}$$

Finally, we use damping to partially update the global approximate site. First, we update the parameters of the global site as follows $\Theta^{new} = (1-\eta)\Theta^{old} + \eta\Theta_i$ where Θ^{old} are the current parameters of the global site, and Θ_i are the parameters for the approximation of a single likelihood. Then, we introduce the global approximate site in the approximate distribution. The learning approach is described in the Algorithm 1.

5 Experimental Results

In this section, we describe the experiments carried out to test the validity of the proposed method on both synthetic and real count data.

Algorithm 1: Stochastic Expectation Propagation (SEP) algorithm for learning a EMSD Mixture model

```
Input : $K$: number of clusters; $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_D\}$: corpus; $p_0(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$: prior
        knowledge
Initialize the approximate site $\tilde{p}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.
If priors are not provided, initialize them to 1 (i.e. $p_0(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$=1)
Compute the approximate distribution $q(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by calculating the average
  effect $\tilde{p}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})^D$ of the likelihood and introducing the priors $p_0$
while not convergence do
    for $x_i$ in $\mathcal{X}$ do
        Compute the cavity distribution $q^{\backslash 1}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by removing the
          contribution of one of the copies of the approximate site.
        Match moments of the tilted distribution $\hat{p}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and approximate
          posterior $q^{new}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by minimizing $D_{KL}(\hat{p} \parallel q^{new})$.
        Compute the parameters of a revised approximate site after matching
          the moments.
        Make a partial update to the approximate site and include the
          approximate site in the approximate distribution.
    end
end
Estimate mixing weights $\pi_j$
```

5.1 Synthetic Dataset

We create a synthetic dataset $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^D$ by using the probabilistic mixture model with $D = 210$ data points. We use $K = 3$ components, each of which is an EMSD distribution where the mixing weights are uniformly sampled. For simplicity, we set a fixed value of 1 for the scale parameter of the Scaled Dirichlet. Since the shape parameter is commonly $\alpha_w \ll 1$ [5], we sample from a Beta distribution.

We initialize the priors of the model with covariance matrix $5\boldsymbol{I}$ and $3\boldsymbol{I}$ for the scale and shape parameter. Random values are used for the prior means and mixing weights parameter. We set a step size of $\eta = 0.1$ and approximate the posterior using SEP. Table 1 show the obtained estimates. The mixing weights can be estimated using the expected value of π_j; for instance, $\mathbb{E}\left[\pi_j\right] = a_j^{'} / \sum_{j=1}^{K} a_j^{'}$.

The used parameters as well as the estimated values are shown in Table 1. We notice that estimates are very close to the target values. Since we need to store only the local and global parameters, we emphasize the fact that SEP reduces memory consumption allowing us to scale EP.

5.2 Sentiment Analysis

We analyze the problem of sentiment analysis in the setting when online users employ online platforms to express opinions or experiences regarding a product or service through reviews. We exploit these data to investigate the validity

Table 1. Original parameters and estimated parameters for the mixture of EMSD using the proposed approach.

j	π	α	β
Real			
1	0.333	$[0.610, 0.318, 0.646]$	**1**
2	0.333	$[0.556, 0.188, 0.848]$	**1**
3	0.334	$[0.129, 0.891, 0.507]$	**1**
Estimation			
1	0.335	$[0.663, 0.305, 0.676]$	$[1.082, 1.055, 1.062]$
2	0.332	$[0.573, 0.098, 0.720]$	$[0.963, 1.027, 0.996]$
3	0.333	$[0.193, 0.858, 0.527]$	$[1.087, 0.976, 1.002]$

of our framework where we know the right number of components (i.e. positive/negative, $K = 2$). We use three benchmark datasets [13,29]: 1) Amazon Review Polarity; 2) Yelp review Polarity; 3) IMDB Movie Reviews. This section presents the details of our experimentation and its results.

Before describing the experimental results, we first outline the key properties of the datasets and the performed setup. We pre-process the dataset as follows: 1) lowercase all text; 2) remove non-alphabetical characters; 3) lemmatize text. All datasets are reviews and contain two labels indicating whether the post has a positive or negative sentiment.

Amazon Review Polarity contains 180k customer reviews that span a period of 18 years, for products on the *Amazon.com* website. The dataset has an average of 75 words per review with a vocabulary size of over 55k unique words.

Yelp Review Polarity contains 560k user reviews from *Yelp* with an average of 133 words with $> 85k$ unique words. The Yelp dataset contains a polarity label by considering stars 1 and 2 negative, and 3 and 4 positive reviews about local businesses.

IMDB movie reviews this dataset consists of 50K movie reviews with an average 231 words per review and a vocabulary size of over 76k unique words. Ratings on IMDB are given as star values $\in [1, 10]$ which were linearly mapped to $[0, 1]$ to use as document labels; negative and positive, respectively.

We compare the clustering performance of EMSD mixture model using the proposed SEP to different models with the same approach and different learning techniques such as Expectation Propagation (EP), and maximum-likelihood (ML) for parameter estimation. More precisely, we compared the performance of EMSD models to the following models that use maximum-likelihood for estimating its parameters. Firstly, we have a mixture of Multinomials (MM) [2]. Even though the MM is appropriate for modeling common words, not words burstiness problem, we add it to the comparison to evaluate its predictive power. Next, we make a comparison with different models that capture the words bustiness problem such as Dirichlet Compound Multinomial (DCM) [14], the Exponential approximation to the Dirichlet Compound Multinomial (EDCM) [5], the Multinomial Scaled Dirichlet (MSD) [25], and the Exponential approximation to the Multinomial Scaled Dirichlet (EMSD) [28]. Furthermore, we compare to the performance of EDCM mixture model in case of considering EP for parameter estimation as we have recently proposed in [23]. We evaluate the performance of the considered models according to precision and recall as illustrated in Table 2.

Table 2. Results on the three text datasets. Comparison using precision and recall. ML: maximum-likelihood; EP: expectation propagation; SEP: sthocastic expectation propagation.

		Dataset		
Metrics		Amazon	Yelp	IMDB
Precision	ML-MM	50.83	89.12	64.18
	ML-DCM	55.65	91.01	71.14
	ML-EDCM	80.65	89.25	78.54
	EP-EDCM	86.91	80.50	86.36
	ML-MSD	82.21	86.96	84.00
	ML-EMSD	83.31	87.23	85.00
	SEP-EMSD (ours)	86.35	82.83	86.83
Recall	ML-MM	51.99	89.20	64.40
	ML-DCM	63.94	91.01	89.45
	ML-EDCM	80.88	89.28	89.33
	EP-EDCM	84.82	93.83	85.94
	ML-MSD	82.21	87.09	84.00
	ML-EMSD	83.57	87.28	86.00
	SEP-EMSD (ours)	83.91	90.02	87.64

In general, most models are superior to the Multinomial mixture model (except for Yelp dataset). We notice that SEP gives comparable results to the EDCM model in terms of precision and recall. Additionally, we evaluate an EDCM mixture that uses EP for parameter learning where we can assume that SEP is computing similar approximations to EP with the advantage that there is no need to store the parameters for each of the approximate sites. One of the main advantages is that we only store the local and global parameters, reducing memory usage. More specifically, for the Amazon dataset, EP and SEP are superior in terms of precision and recall compared with most models that use maximum-likelihood estimation. Our intuition is that the length of documents plays a critical role in parameter estimation. That is, in the Amazon dataset, for example, we obtain better precision and recall using a Bayesian approach given that the document length is relatively shorter than in the other two datasets.

6 Conclusions

In this paper, we propose a Stochastic Expectation Propagation (SEP) algorithm to learn a finite EMSD mixture model. We derive the mathematical framework using SEP, and since performing moment matching leads to an intractable integral, we use sampling in order to compute its moments. Then, we evaluate the proposed approach on both synthetic and real data and notice that SEP-EMSD provides comparable results to traditional approaches and in some cases is superior. Although we evaluated the proposed learning method with text data, we can use any type of count data such as a clustering of visual words for images or videos. It is noticeable that SEP does not need a site per data point and similar to variational inference maintains a global posterior approximation that is updated locally and reduces memory consumption.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
2. Bouguila, N., Ziou, D.: Unsupervised learning of a finite discrete mixture model based on the multinomial dirichlet distribution: Application to texture modeling. In: Fred, A.L.N. (ed.) Pattern Recognition in Information Systems, Proceedings of the 4th International Workshop on Pattern Recognition in Information Systems, PRIS 2004, in conjunction with ICEIS 2004, Porto, Portugal, April 2004, pp. 118–127. INSTICC Press (2004)
3. Boyd-Graber, J., Hu, Y., Mimno, D., et al.: Applications of topic models. Found. Trends® Inf. Retrieval **11**(2–3), 143–296 (2017)
4. Bui, T.D., Hernández-Lobato, J.M., Li, Y., Hernández-Lobato, D., Turner, R.E.: Training deep gaussian processes using stochastic expectation propagation and probabilistic backpropagation. arXiv preprint arXiv:1511.03405 (2015)
5. Elkan, C.: Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 289–296. ACM (2006)
6. Fan, W., Bouguila, N.: Expectation propagation learning of a dirichlet process mixture of beta-liouville distributions for proportional data clustering. Eng. Appl. Artif. Intell. **43**, 1–14 (2015)
7. Gelman, A., et al.: Expectation propagation as a way of life: a framework for bayesian inference on partitioned data. arXiv preprint arXiv:1412.4869 (2017)
8. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. J. Mach. Learn. Res. **14**(1), 1303–1347 (2013)
9. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
10. Li, Y., Hernández-Lobato, J.M., Turner, R.E.: Stochastic expectation propagation. In: Advances in Neural Information Processing Systems, pp. 2323–2331 (2015)
11. Lochner, R.H.: A generalized dirichlet distribution in bayesian life testing. J. Roy. Stat. Soc. Ser. B (Methodol.) **37**(1), 103–113 (1975)
12. Ma, Z., Leijon, A.: Expectation propagation for estimating the parameters of the beta distribution. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2082–2085. IEEE (2010)

13. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 142–150. Association for Computational Linguistics (2011)
14. Madsen, R.E., Kauchak, D., Elkan, C.: Modeling word burstiness using the dirichlet distribution. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 545–552. ACM (2005)
15. Margaritis, D., Thrun, S.: A bayesian multiresolution independence test for continuous variables. arXiv preprint arXiv:1301.2292 (2013)
16. Minka, T.: Estimating a dirichlet distribution (2000)
17. Minka, T.: Power ep. Technical report, Microsoft Research, Cambridge (2004)
18. Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, pp. 352–359. Morgan Kaufmann Publishers Inc. (2002)
19. Minka, T.P.: Expectation propagation for approximate bayesian inference. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pp. 362–369. Morgan Kaufmann Publishers Inc. (2001)
20. Minka, T.P.: A family of algorithms for approximate Bayesian inference. Ph.D. thesis, Massachusetts Institute of Technology (2001)
21. Neal, R.M.: Probabilistic inference using markov chain monte carlo methods (1993)
22. Opper, M., Winther, O.: A bayesian approach to on-line learning. In: On-line Learning in Neural Networks, pp. 363–378 (1998)
23. Sumba, X., Zamzami, N., Bouguila, B.: Improving the edcm mixture model with expectation propagation. In: 2020 Association for the Advancement of Artificial Intelligence AAAI. FLAIRS 33 (2020)
24. Wong, T.T.: Alternative prior assumptions for improving the performance of naïve bayesian classifiers. Data Min. Knowl. Disc. **18**(2), 183–213 (2009)
25. Zamzami, N., Bouguila, N.: Text modeling using multinomial scaled dirichlet distributions. In: Mouhoub, M., Sadaoui, S., Ait Mohamed, O., Ali, M. (eds.) IEA/AIE 2018. LNCS (LNAI), vol. 10868, pp. 69–80. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92058-0_7
26. Zamzami, N., Bouguila, N.: An accurate evaluation of msd log-likelihood and its application in human action recognition. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1–5. IEEE (2019)
27. Zamzami, N., Bouguila, N.: Hybrid generative discriminative approaches based on multinomial scaled dirichlet mixture models. Appl. Intell **49**(11), 3783–3800 (2019)
28. Zamzami, N., Bouguila, N.: A novel scaled dirichlet-based statistical framework for count data modeling: unsupervised learning and exponential approximation. Pattern Recogn. **95**, 36–47 (2019)
29. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems, pp. 649–657 (2015)