EP-Based Infinite Inverted Dirichlet Mixture Learning: Application to Image Spam Detection

Wentao Fan[1], Sami Bourouis[2], Nizar Bouguila[3(✉)], Fahd Aldosari[4], Hassen Sallay[4], and K. M. Jamil Khayyat[4]

[1] Huaqiao University, Xiamen, China
fwt@hqu.edu.cn
[2] Taif university, Taif, Kingdom of Saudi Arabia
s.bourouis@tu.edu.sa
[3] Concordia University, Montreal, QC, Canada
nizar.bouguila@concordia.ca
[4] Umm Al-Qura University, Makkah, Saudi Arabia
{fmdosari,hmsallay,kmkhayyat}@uqu.edu.sa

Abstract. We propose in this paper a new fully unsupervised model based on a Dirichlet process prior and the inverted Dirichlet distribution that allows the automatic inferring of clusters from data. The main idea is to let the number of mixture components increases as new vectors arrive. This allows answering the model selection problem in a elegant way since the resulting model can be viewed as an infinite inverted Dirichlet mixture. An expectation propagation (EP) inference methodology is developed to learn this model by obtaining a full posterior distribution on its parameters. We validate the model on a challenging application namely image spam filtering to show the merits of the framework.

1 Introduction

Contemporary times have witnessed an exponential increase in the volume of data generated everyday. These data can be textual or visual (in the form of images and videos). The organization, analysis, and modeling of these data is a crucial problem that has been growing in importance. One of the challenging data analysis tasks is clustering. Finite mixture models have been widely used for clustering since they offer a formal approach for unsupervised learning [1]. Many statistical frameworks based on finite mixture models have been proposed in the past. Despite the fact that the majority of these frameworks assume that the per-components densities are Gaussian, some recent works have considered other densities by taking the nature of the data into account. Examples of these research works includes inverted Dirichlet-based models, which we will consider in this paper, for semi-bounded data (i.e. positive vectors) [2]. An important problem when deploying mixture models is the automatic selection of the number of components. This problem has been tackled in [2] using the minimum

M. Mouhoub et al. (Eds.): IEA/AIE 2018, LNAI 10868, pp. 342–354, 2018.
https://doi.org/10.1007/978-3-319-92058-0_33

message length approach that has been shown in [3] to be a generalization of several other well known model's selection criteria. The main problem with this approach is that it needs running the estimation algorithm for different number of components and then selecting the optimal one according to the resulting message length which is actually time consuming. In this paper we go a step further by considering an infinite number of mixture components components [4] via a nonparametric Bayesian approach [5].

Nonparametric Bayesian approaches, which are statistically well based, have received a lot of attention recently and are now well-understood and accepted. These approaches are generally based on considering Dirichlet processes (DPs) [6]. DPs allows a technically sound approach for unsupervised Bayesian clustering and have been deployed in a variety of domains and applications such as computer vision, pattern recognition, data mining, and information retrieval [7–9]. In this paper, we rely on DPs to develop our nonparametric Bayesian model. Indeed, the proposed work can be viewed as an extension of the finite mixture framework developed in [10], based on the inverted Dirichlet [2], to the infinite case. The main idea is to allow the complexity and accuracy of the model to increase as the data size increases [11–13]. Having the infinite inverted Dirichlet mixture model in hand, a challenging problem that we will tackle in this paper is the learning of its parameters. Markov Chain Monte Carlo (MCMC) techniques have dominated the literature in the case of infinite mixture models learning. Unfortunately MCMC approaches have been shown to be computationally extensive. Thus, we consider here, a deterministic approximation technique to MCMC, known as expectation propagation (EP), has been introduced and has been shown to be a good learning alternative [14–16]. EP is an extension to assumed-density filtering (ADF) [17] which is a one pass, sequential approximation method. In contrast to the ADF, the order of the input data points is not crucial in the EP inference and its inference accuracy is improved by reusing the data points many times. This allows the simultaneous estimation of the parameters and selection of the number of clusters. The resulting statistical framework is applied to tackle the challenging problem of image spam detection where visual content of emails is considered for the filtering task.

The remainder of this paper is organized as follows. Section 2 presents our statistical framework. Section 3 develops in details the model's expectation propagation learning approach. The experimental evaluation is given in Sect. 4. Finally, Sect. 5 draws the conclusion.

2 Model Specification

In this work, we focus on the Dirichlet process mixture of inverted Dirichlet distributions, which can also be considered as an infinite inverted Dirichlet mixture model since it is composed of an infinite number of mixture components.

2.1 Finite Inverted Dirichlet Mixture Model

If a D-dimensional random positive vector $\boldsymbol{X} = (X_1, \ldots, X_D)$ is distributed according to the inverted Dirichlet mixture model with J components, then the probability density function of $\boldsymbol{X}$ is given by

$$p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\alpha}) = \sum_{j=1}^{J} \pi_j \mathcal{ID}(\boldsymbol{X}|\boldsymbol{\alpha}_j) \ , \tag{1}$$

where $\boldsymbol{\pi} = \{\pi_j\}$ are the mixing proportions that have to be positive and sum to unity. $\mathcal{ID}(\boldsymbol{X}|\boldsymbol{\alpha}_j)$ is the inverted Dirichlet distribution associated with the jth component and is parameterized by $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jD})$ as

$$\mathcal{ID}(\boldsymbol{X}_i|\boldsymbol{\alpha}_j) = \frac{\Gamma(\sum_{l=1}^{D+1} \alpha_{jl})}{\prod_{l=1}^{D+1} \Gamma(\alpha_{jl})} \prod_{l=1}^{D} X_{il}^{\alpha_{jl}-1} \big(1 + \sum_{l=1}^{D} X_{il}\big)^{-\sum_{l=1}^{D+1} \alpha_{jl}} \tag{2}$$

where $0 < X_{il} < \infty$ for $l = 1, \ldots, D$, $\alpha_{jl} > 0$ for $l = 1, \ldots, D+1$. The mean, variance and covariance of the inverted Dirichlet distribution are given by

$$E(X_l) = \frac{\alpha_l}{(\alpha_{D+1} - 1)} \tag{3}$$

$$var(X_l) = \frac{\alpha_l(\alpha_l + \alpha_{D+1} - 1)}{(\alpha_{D+1} - 1)^2(\alpha_{D+1} - 2)} \tag{4}$$

$$cov(X_a, X_b) = \frac{\alpha_a \alpha_b}{(\alpha_{D+1} - 1)^2(\alpha_{D+1} - 2)} \tag{5}$$

2.2 Stick-Breaking Representation

In this section, we extend the finite inverted Dirichlet mixture model to the infinite counterpart using Dirichlet process prior with stick-breaking construction [18]. Assume that a random distribution G is distributed according to a Dirichlet process $G \sim \mathrm{DP}(b, H)$ with the base distribution H and concentration parameter b, its stick-breaking representation can be described as

$$\lambda_j \sim \mathrm{Beta}(1, b), \qquad \theta_j \sim H, \qquad \pi_j = \lambda_j \prod_{s=1}^{j-1}(1 - \lambda_s), \qquad G = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j} \tag{6}$$

where δ_{θ_j} is the Dirac delta measure centered at θ_j, π_j are the mixing proportions with the constraint that $\sum_{j=1}^{\infty} \pi_j = 1$. If a set of N i.i.d observations $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ follows an inverted Dirichlet mixture model with an infinite number of components, then the probability of $\mathcal{X}$ is defined as

$$p(\mathcal{X}|\boldsymbol{\pi}, \boldsymbol{\alpha}) = \prod_{i=1}^{N} \big[\sum_{j=1}^{\infty} \pi_j \mathcal{ID}(\boldsymbol{X}_i|\boldsymbol{\alpha}_j)\big] \tag{7}$$

3 EP-Based Learning

3.1 Expectation Propagation

In this subsection, a brief introduction to the EP approximation scheme is presented. Consider an observed data set of N i.i.d vectors $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ which follow a model with unknown parameter Θ, then the joint distribution of $\mathcal{X}$ and Θ can be represented in the form of a product of factors as in [14]:

$$p(\mathcal{X}, \Theta) = \prod_i f_i(\Theta) \tag{8}$$

One possible factorization is to consider $N+1$ terms (the prior term + the data terms) as: $f_0(\Theta) = p(\Theta)$ and $f_i(\Theta) = p(\boldsymbol{X}_i|\Theta)$, $i = 1, \ldots, N$. The main idea of the EP algorithm is to approximate the posterior distribution $p(\Theta|\mathcal{X})$ by a product of factors:

$$q(\Theta) = \frac{\prod_i \widetilde{f}_i(\Theta)}{\int \prod_i \widetilde{f}_i(\Theta) d\Theta} \tag{9}$$

where each factor $\widetilde{f}_i(\Theta)$ is an appropriate approximation to $f_i(\Theta)$. In the EP learning framework, we first initialize all the factors $\widetilde{f}_i(\Theta)$, then each factor is optimized sequentially in the context of the remaining factors. In order to estimate a specific factor $\widetilde{f}_j(\Theta)$, we first remove it from the current approximation to the posterior as

$$q^{\setminus j}(\Theta) = \frac{q(\Theta)}{\widetilde{f}_j(\Theta)} \tag{10}$$

We then obtain a new distribution by combining Eq. (10) with the true factor $f_j(\Theta)$ as

$$\widehat{p}(\Theta) = \frac{f_j(\Theta) q^{\setminus j}(\Theta)}{\int f_j(\Theta) q^{\setminus j}(\Theta) d\Theta} \tag{11}$$

Consequently, we can update the approximated posterior $q(\Theta)$ by minimizing the KL divergence: $\mathrm{KL}\big(\widehat{p}(\Theta) \parallel q(\Theta)\big)$. This is achieved by matching the sufficient statistics of $q(\Theta)$ to the corresponding moments of $\widehat{p}(\Theta)$. Then, we can update the approximating factor $\widetilde{f}_j(\Theta)$ as

$$\widetilde{f}_j(\Theta) = Z_j \frac{q(\Theta)}{q^{\setminus j}(\Theta)} \tag{12}$$

where $Z_j = \int f_j(\Theta) q^{\setminus j}(\Theta) d\Theta$ is a normalization constant. Therefore, each factor can be updated iteratively in the context of remaining factors as described in the above steps until convergence.

3.2 EP Model Learning

In this section, we adopt Expectation Propagation (EP) to learn the infinite inverted Dirichlet mixture model. Since our model is fully Bayesian, we need to introduce priors for parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$. Based on the stick-breaking representation of the Dirichlet process as defined in Eq.(6), the prior of $\boldsymbol{\lambda}$ is a particular Beta distribution parameterized by $a_j = 1$ and b_j in the form of

$$p(\boldsymbol{\lambda}) = \prod_{j=1}^{\infty} \text{Beta}(\lambda_j | a_j, b_j) = \prod_{j=1}^{\infty} b_j (1-\lambda_j)^{b_j - 1} \tag{13}$$

For the parameter $\boldsymbol{\alpha}_j$ of the jth component of the inverted Dirichlet mixture model, inspired by [19] in which a Gaussian assumption is adopted as the priors for the parameters of Beta distribution, we adopt a $D+1$ dimensional Gaussian with mean vector $\boldsymbol{\mu}_j$ and the precision matrix A_j (the inverse covariance matrix) as the prior for $\boldsymbol{\alpha}_j$ as:

$$p(\boldsymbol{\alpha}_j) = \mathcal{N}(\boldsymbol{\alpha}_j | \boldsymbol{\mu}_j, A_j) = \frac{|A_j|^{1/2}}{(2\pi)^{(D+1)/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\alpha}_j - \boldsymbol{\mu}_j)^T A_j (\boldsymbol{\alpha}_j - \boldsymbol{\mu}_j) \right) \tag{14}$$

The first step of EP learning is to initialize all of the approximating factors $\widetilde{f}_i(\Theta)$, where $\Theta = \{\boldsymbol{\alpha}, \boldsymbol{\lambda}\}$, by initializing all the involved hyperparameters: $\{a_j, b_j, \boldsymbol{\mu}_j, A_j\}$. We also truncate the stick-breaking representation for the infinite inverted Dirichlet mixture model at a value of J as:

$$\lambda_J = 1\ , \qquad \pi_j = 0 \text{ when } j > J\ , \qquad \sum_{j=1}^{J} \pi_j = 1 \tag{15}$$

where the truncation level J is infered automatically during the EP learning process. Next, the posterior approximation $q(\Theta)$ is initialized by setting $q(\Theta) \propto \prod_i \widetilde{f}_i(\Theta)$, where the corresponding hyperparameters of $q(\Theta)$ are denoted as $\{a_j^*, b_j^*, \boldsymbol{\mu}_j^*, A_j^*\}$. Since each approximated term $\widetilde{f}_i(\Theta)$ is in exponential form, we can easily compute the hyperparameters of $q(\Theta)$ according to [14,16],

$$a_j^* = \sum_i a_{i,j} - N, \qquad b_j^* = \sum_i b_{i,j} - N \tag{16}$$

$$\boldsymbol{\mu}_j^* = \big(\sum_i A_{i,j}^{-1}\big)\big(\sum_i A_{i,j}\boldsymbol{\mu}_{i,j}\big), \qquad A_j^* = \sum_i A_{i,j} \tag{17}$$

In order to update the factor $\widetilde{f}_i(\Theta)$, we have to remove it from the posterior $q(\Theta)$. Then, the corresponding hyperparameters can be computed analytically as

$$a_j^{\backslash i} = a_j^* - a_{i,j} + 1, \qquad b_j^{\backslash i} = b_j^* - b_{i,j} + 1 \tag{18}$$

$$\boldsymbol{\mu}_j^{\backslash i} = \big(A_j^{\backslash i}\big)^{-1}\big(A_j^*\boldsymbol{\mu}_j^* - A_{i,j}\boldsymbol{\mu}_{i,j}\big), \qquad A_j^{\backslash i} = A_j^* - A_{i,j} \tag{19}$$

Next, the updated posterior $\widehat{p}(\Theta)$ can be calculated as

$$\widehat{p}(\Theta) = \frac{1}{Z_i} f_i(\Theta) q^{\setminus i}(\Theta) \tag{20}$$

where the normalization constant Z_i is evaluated by

$$Z_i = \int f_i(\Theta) q^{\setminus i}(\Theta) \tag{21}$$

It is noteworthy that the normalization constant Z_i in Eq. (20) is analytically intractable, since it involves an integration over the product of an inverted Dirichlet and a Gaussian distribution. To tackle this problem, we apply the Laplace approximation to approximate the integrand with a Gaussian distribution as suggested in [19] (details can be viewed in Appendix A). After obtaining Z_i and $\widehat{p}(\Theta)$, we can revise the posterior $q(\Theta)$ by matching its sufficient statistics to the corresponding moments of $\widehat{p}(\Theta)$. This is achieved by calculating the partial derivative of $\ln Z_i$ with respect to the corresponding model hyperparameters. For $a_j^{\setminus i}$, we can calculate the partial derivative as

$$\begin{aligned} \nabla_{a_j}^{\setminus i} \ln Z_i &= \frac{1}{Z_i} \int f_i(\Theta) \frac{q^{\setminus i}(\Theta)}{q^{\setminus i}(\lambda_j^{\setminus i})} \frac{\partial}{\partial a_j^{\setminus i}} q^{\setminus i}(\lambda_j^{\setminus i}) d\Theta \\ &= E_{\widehat{p}}[\ln \lambda_j] + \Psi(a_j^{\setminus i} + b_j^{\setminus i}) - \Psi(a_j^{\setminus i}) \end{aligned} \tag{22}$$

By applying moment matching, we obtain

$$E_{\widehat{p}}[\ln \lambda_j] = E_q[\ln \lambda_j] = \Psi(a_j^*) - \Psi(a_j^* + b_j^*) \tag{23}$$

Similarly, we can compute the partial derivatives of $\ln Z_i$ with respect to the other model hyperparameters:

$$\nabla_{b_j}^{\setminus i} \ln Z_i = E_{\widehat{p}}[1 - \ln \lambda_j] + \Psi(a_j^{\setminus i} + b_j^{\setminus i}) - \Psi(b_j^{\setminus i}) \tag{24}$$

$$\nabla_{\boldsymbol{\mu}_j}^{\setminus i} \ln Z_i = A_j^{\setminus i} E_{\widehat{p}}[\boldsymbol{\alpha}_j] - A_j^{\setminus i} \boldsymbol{\mu}_j^{\setminus i} \tag{25}$$

$$\nabla_{A_j}^{\setminus i} \ln Z_i = \frac{1}{2} \left\{ |(A_j^{\setminus i})^{-1}| - \left[\sum_{l=1}^{D+1} E_{\widehat{p}}[\alpha_{jl}^2] - 2E_{\widehat{p}}[\alpha_{jl}] \mu_{jl}^{\setminus i} + (\mu_{jl}^{\setminus i})^2 \right] \right\} \tag{26}$$

The right hand sides in the above equations can be computed analytically by using Eq. (39) in the Appendix. Furthermore, the expectations in the above equations can be acquired by applying the moment matching technique as

$$E_{\widehat{p}}[\boldsymbol{\alpha}_j] = E_q[\boldsymbol{\alpha}_j] = \boldsymbol{\mu}_j^*, \qquad E_{\widehat{p}}[\boldsymbol{\alpha}_j^2] = E_q[\boldsymbol{\alpha}_j^2] = (\boldsymbol{\mu}_j^*)^2 \tag{27}$$

By substituting the above expectations into the corresponding partial derivative equations, we can update the hyperparameters of $q(\Theta)$. After obtaining $q(\Theta)$ and $q^{\backslash i}(\Theta)$, we can update the revised hyperparameters for the approximating factor f_i as

$$a_{i,j} = a_j^* - a_j^{\backslash i} + 1, \qquad b_{i,j} = b_j^* - b_j^{\backslash i} + 1 \tag{28}$$

$$\boldsymbol{\mu}_{i,j} = A_{i,j}^{-1}(A_j^* \boldsymbol{\mu}_j^* - A_j^{\backslash i} \boldsymbol{\mu}_j^{\backslash i}), \qquad A_{i,j} = A_j^* - A_j^{\backslash i} \tag{29}$$

The above procedure is repeated until the hyperparameters of the approximating factor converge. The same procedure is applied sequentially for the remaining factors. The complete learning process is summarized in Algorithm 1.

Algorithm 1. EP learning of infinite inverted Dirichlet mixture

1: Choose the initial truncation level J.
2: Initialize the approximating factors $\widetilde{f}_i(\Theta)$ by initializing all the involved hyperparameters $\{a_j, b_j, \boldsymbol{\mu}_j, A_j\}$.
3: Initialize the posterior approximation by setting $q(\Theta) \propto \prod_i \widetilde{f}_i(\Theta)$. The hyperparameters of $q(\Theta)$ are calculated by Eqs. (16) and (17).
4: **repeat**
5: Select a factor $\widetilde{f}_i(\Theta)$ to refine.
6: Remove $\widetilde{f}_i(\Theta)$ from the posterior $q(\Theta)$ by division $q^{\backslash i}(\Theta) = q(\Theta)/\widetilde{f}_i(\Theta)$.
7: Evaluate the new posterior by setting the sufficient statistics (moments) of $q(\Theta)$ to the corresponding moments of $\widehat{p}(\Theta)$.
8: Update the factor $\widetilde{f}_i(\Theta)$ by updating the corresponding hyperparameters as in Eqs. (28) and (29).
9: **until** Convergence criterion is reached.
10: Calculate the expected value of λ_j as $E[\lambda_j] = \frac{a_j^*}{a_j^* + b_j^*}$, and submit it into Eq. (6) to obtain the estimated values of the mixing coefficients π_j.
11: Detect the optimal number of components J by eliminating the components with small mixing coefficients close to 0.

4 Experimental Results: Image Spam Detection

Exchanging emails play an important role in our daily activities. Spam filtering has been one of the most challenging problems in digital communication in the last couple of decades [20]. Spams do not only compromise ressources, but also cause security problems. Various techniques and approaches have been proposed in the past to deal with this problem. Many of the proposed solutions are based on machine learning techniques. These techniques are able to detect and extract hidden patterns that could discriminate between legitimates and spam emails. Two major challenges in the spam filtering problem are its dynamic nature and dealing with the non-textual content. Thus, a good filtering approach should be adaptive [21,22] and should be able to deal with the presence in images in emails. Contrary to textual data, images pose several significant challenges. Indeed, it

is important to choose an appropriate representation that can describe well the content of the image present in a given email. And the resulting representation should provide a semantically meaningful output which is very difficult taking into account the fact that an image possesses a rich structure. Embedding spam images into emails is a successful trick that is widely used now by spammers. This trick is generally referred to as image-based spam and has drawn some attention recently [23–26]. In this section, we validate our proposed model in the challenging task of image spam detection. Our model is considered simultaneously with the probabilistic Latent Semantic Analysis (pLSA) model [27] with bag-of-words representation [28]. We have considered three challenging spam data sets of images extracted from real spam in our experiments: the personal spam emails collected by Dredze et al. [29], a subset of the publicly available SpamArchive corpus used by [24,29] and the Princeton spam image benchmark[1]. We have used one common ham data set of images which was collected and used by Dredze et al. [29]. In total, there are 2,550 images in the Dredze ham data set, 3,210 images in the Dredze spam data set, 3,550 images in the SpamArchive and 1,071 images in Princeton spam image benchmark. Sample spam images are shown in Fig. 1. We have downsampled all images to the spatial resolution of 100×100 pixels as a preprocessing step. In our experiments, we have randomly divided each data set (both ham and spam) into two halves: one for constructing the visual vocabulary and another for testing.

Fig. 1. Sample spam images from Dredze spam data set.

We have proceeded as following in order to construct the visual vocabulary, First, the key points of each image were detected using the Difference-of-Gaussian (DoG) interest point detector and described using Scale-Invariant Feature Transform (SIFT) resulting in a 128-dimensional vector for each key point [30]. Then, the K-Means algorithm was used to cluster all the SIFT vectors into a visual vocabulary. We have constructed the visual vocabulary by setting the number of clusters (i.e. number of visual words) to 800, 1000 and 850, respectively for each data set. The pLSA model was applied by considering 45 aspects for all data sets and each image in the data set was then represented

[1] http://www.cs.jhu.edu/~mdredze/datasets/image_spam.

by a 45-dimensional vector. Finally, the resulting vectors were clustered by our mixture model. The entire procedure was repeated 10 times for evaluating the performance of our approach. Tables 1, 2 and 3 represent the average confusion matrices for detecting spam images of each data set using our mixture model. In these tables, SI stands for spam images while HI denotes ham images. Moreover, we compared the performances of our expectation propagation infinite inverted Dirichlet mixture (EPInInDM), infinite inverted Dirichlet mixture (InInDM) learned via MCMC technique as proposed in [31], variational infinite inverted Dirichlet mixture model (varInInDM) as proposed in [32], and variatioanl infinite Gaussian mixture model (varInGM) model in terms of the average classification accuracy rate and the average false positive rate. The corresponding results as well as the number of correct detected images (both spam and ham) are illustrated in Table 4. Based on this table, the proposed EPInInDM model obtains the highest average accuracy rate and lowest false positive rate.

Table 1. Dredze.

	SI	HI
SI	1250	355
HI	60	1215

Table 2. SpamArchive.

	SI	HI
SI	1351	424
HI	93	1182

Table 3. Princeton.

	SI	HI
SI	378	158
HI	61	1214

Table 4. The number of correct detected images (both spam and ham) $\hat{N}$, and the average classification accuracy rate (Acc.) of image spam detection computed using different algorithms over 10 random runs.

	Dredze		SpamArchive		Princeton	
	$\hat{N}$	Acc. (%)	$\hat{N}$	Acc. (%)	$\hat{N}$	Acc. (%)
EPInInDM	2465	85.59	2533	83.04	1592	87.90
InInDM	2401	83.36	2488	81.49	1530	84.39
varInInDM	2290	79.51	2426	79.53	1441	79.50
varInGM	2263	78.58	2360	77.39	1435	79.25

5 Conclusion

We proposed an infinite mixture model based on inverted Dirichlet distribution to model positive vectors. We developed and evaluated an expectation propagation algorithm to learn the parameters of the proposed infinite model. The experiments confirm the power of the proposed approach in tackling a very challenging problem namely image spam filtering. The presented statistical framework is of general applicability, as it can work in any situation where positive feature vectors are extracted.

Acknowledgements. The authors would like to thank the Deanship of Scientific Research at umm Al-Qura University for the continuous support. This work was supported financially by the Deanship of Scientific Research at Umm Al-Qura University under the grant number 15-COM-3-1-0006. The first author was supported by the National Natural Science Foundation of China (61502183).

A The calculation of Z_i in Eq. (21)

The normalized constant Z_i in Eq. (21) can be calculated as

$$Z_i = \int f_i(\Theta) q^{\setminus i}(\Theta) d\Theta = \sum_{j=1}^{J} \bar{\lambda}_j \prod_{s=1}^{j-1} (1 - \bar{\lambda}_s) \int \mathcal{ID}(\boldsymbol{X}_i|\boldsymbol{\alpha}_j) N(\boldsymbol{\alpha}_j|\boldsymbol{\mu}_j^{\setminus i}, A_j^{\setminus i}) \mathrm{d}\boldsymbol{\alpha}_j \tag{30}$$

where $\bar{\lambda}_j$ is the expected value of λ_j. Since the integration involved in Eq. (30) is analytically intractable, we tackle this problem by adopting the Laplace approximation to approximate the integrand with a Gaussian distribution [19]. First, we define $h(\boldsymbol{\alpha}_j)$ as the integrand in Eq. (30):

$$h(\boldsymbol{\alpha}_j) = \mathcal{ID}(\boldsymbol{X}_i|\boldsymbol{\alpha}_j) \mathcal{N}(\boldsymbol{\alpha}_j|\boldsymbol{\mu}_j^{\setminus i}, A_j^{\setminus i}) \tag{31}$$

Then, the normalized distribution for this integrand which is indeed a product of a Dirichlet distribution and a Gaussian distribution is given by

$$\mathcal{H}(\boldsymbol{\alpha}_j) = \frac{h(\boldsymbol{\alpha}_j)}{\int h(\boldsymbol{\alpha}_j) d\boldsymbol{\alpha}_j} \tag{32}$$

Our goal for the Laplace method is to find a Gaussian approximation which is centered on the mode of the distribution $\mathcal{H}(\boldsymbol{\alpha}_j)$. We may obtain the mode $\boldsymbol{\alpha}_j^*$ numerically by setting the first derivative of $\ln h(\boldsymbol{\alpha}_j)$ to 0, where

$$\begin{aligned} \ln h(\boldsymbol{\alpha}_j) = & \ln \frac{\Gamma(\sum_{l=1}^{D+1} \alpha_{jl})}{\prod_{l=1}^{D+1} \Gamma(\alpha_{jl})} + \sum_{l=1}^{D} (\alpha_{jl} - 1) \ln X_{il} - \sum_{l=1}^{D+1} \alpha_{jl} \\ & \ln(1 + \sum_{l=1}^{D} X_{il}) - \frac{1}{2} (\boldsymbol{\alpha}_j - \boldsymbol{\mu}_j^{\setminus i})^T A_j^{\setminus i} (\boldsymbol{\alpha}_j - \boldsymbol{\mu}_j^{\setminus i}) + \text{const.} \end{aligned} \tag{33}$$

We can calculate the first and second derivatives with respect to $\boldsymbol{\alpha}_j$ as

$$\frac{\partial \ln h(\boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}_j} = \begin{bmatrix} \Psi(\sum_{l=1}^{D+1} \alpha_{jl}) - \Psi(\alpha_{j1}) + \ln X_{i1} - \ln(1 + \sum_{l=1}^{D} X_{il}) \\ \vdots \\ \Psi(\sum_{l=1}^{D+1} \alpha_{jl}) - \Psi(\alpha_{jD}) + \ln X_{iD} - \ln(1 + \sum_{l=1}^{D} X_{il}) \end{bmatrix} - A_j^{\setminus i} (\boldsymbol{\alpha}_j - \boldsymbol{\mu}_j^{\setminus i}) \tag{34}$$

$$\frac{\partial^2 \ln h(\boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}_j^2} = \begin{bmatrix} \Psi'(\sum_{l=1}^{D} \alpha_{jl}) - \Psi'(\alpha_{j1}) & \cdots & \Psi'(\sum_{l=1}^{D} \alpha_{jl}) \\ \vdots & \ddots & \vdots \\ \Psi'(\sum_{l=1}^{D} \alpha_{jl}) & \cdots & \Psi'(\sum_{l=1}^{D} \alpha_{jl}) - \Psi'(\alpha_{jD}) \end{bmatrix} - A_j^{\setminus i} \tag{35}$$

where $\Psi(\cdot)$ is the digamma function. Then, we can approximate $h(\boldsymbol{\alpha}_j)$

$$h(\boldsymbol{\alpha}_j) \simeq h(\boldsymbol{\alpha}_j^*) \exp\left(-\frac{1}{2}(\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_j^*)\widehat{A}_j(\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_j^*)\right) \tag{36}$$

where the precision matrix $\widehat{A}_j$ is given by

$$\widehat{A}_j = - \left. \frac{\partial^2 \ln h(\boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}_j^2} \right|_{\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j^*} \tag{37}$$

Therefore, the integration of $h(\boldsymbol{\alpha}_j)$ can be approximated by using Eq. (36) as

$$\int h(\boldsymbol{\alpha}_j) d\boldsymbol{\alpha}_j \simeq h(\boldsymbol{\alpha}_j^*) \int \exp(-\frac{1}{2}(\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_j^*)\widehat{A}_j(\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_j^*)) d\boldsymbol{\alpha}_j = h(\boldsymbol{\alpha}_j^*) \frac{(2\pi)^{(D+1)/2}}{|\widehat{A}_j|^{1/2}} \tag{38}$$

Finally, we can rewrite Eq. (30) as following:

$$Z_i = \sum_{j=1}^{J} \bar{\lambda}_j \prod_{s=1}^{j-1} (1 - \bar{\lambda}_s) h(\boldsymbol{\alpha}_j^*) \frac{(2\pi)^{(D+1)/2}}{|\widehat{A}_j|^{1/2}} \tag{39}$$

References

1. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
2. Bdiri, T., Bouguila, N.: Positive vectors clustering using inverted dirichlet finite mixture models. Expert Syst. Appl. **39**(2), 1869–1882 (2012)
3. Bouguila, N., Ziou, D.: High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1716–1731 (2007)
4. Rasmussen, C.E.: The infinite Gaussian mixture model. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 554–560. MIT Press (2000)
5. Blackwell, D., MacQueen, J.: Ferguson distributions via pólya urn schemes. Ann. Stat. **1**(2), 353–355 (1973)
6. Korwar, R.M., Hollander, M.: Contributions to the theory of Dirichlet processes. Ann. Prob. **1**, 705–711 (1973)
7. Blei, D.M., Jordan, M.I.: Variational inference for Dirichlet process mixtures. Bayesian Anal. **1**, 121–144 (2005)
8. Bouguila, N., Ziou, D.: A Dirichlet process mixture of Dirichlet distributions for classification and prediction. In: Proceedings of the IEEE Workshop on Machine Learning for Signal Processing (MLSP), pp. 297–302 (2008)
9. Zhang, X., Chen, B., Liu, H., Zuo, L., Feng, B.: Infinite max-margin factor analysis via data augmentation. Pattern Recogn. **52**(Suppl. C), 17–32 (2016)
10. Bertrand, A., Al-Osaimi, F.R., Bouguila, N.: View-based 3D objects recognition with expectation propagation learning. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Porikli, F., Skaff, S., Entezari, A., Min, J., Iwai, D., Sadagic, A., Scheidegger, C., Isenberg, T. (eds.) ISVC 2016. LNCS, vol. 10073, pp. 359–369. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50832-0_35

11. Minka, T., Ghahramani, Z.: Expectation propagation for infinite mixtures. In: NIPS 2003 Workshop on Nonparametric Bayesian Methods and Infinite Models (2003)
12. Bouguila, N.: Infinite Liouville mixture models with application to text and texture categorization. Pattern Recogn. Lett. **33**(2), 103–110 (2012)
13. Bouguila, N., Ziou, D.: A Dirichlet process mixture of generalized Dirichlet distributions for proportional data modeling. IEEE Trans. Neural Netw. **21**(1), 107–122 (2010)
14. Minka, T.: Expectation propagation for approximate Bayesian inference. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI), pp. 362–369 (2001)
15. Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI), pp. 352–359 (2002)
16. Chang, S., Dasgupta, N., Carin, L.: A Bayesian approach to unsupervised feature selection and density estimation using expectation propagation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1043–1050 (2005)
17. Maybeck, P.S.: Stochastic Models, Estimation and Control. Academic Press, New York (1982)
18. Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**, 161–173 (2001)
19. Ma, Z., Leijon, A.: Expectation propagation for estimating the parameters of the beta distribution. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2082–2085 (2010)
20. Zhang, Y., Wang, S., Phillips, P., Ji, G.: Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowl. Based Syst. **64**, 22–31 (2014)
21. Amayri, O., Bouguila, N.: Improved online support vector machines spam filtering using string kernels. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 621–628. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10268-4_73
22. Amayri, O., Bouguila, N.: Online spam filtering using support vector machines. In: Proceedings of the 14th IEEE Symposium on Computers and Communications (ISCC 2009), 5–8 July Sousse, Tunisia, pp. 337–340. IEEE Computer Society (2009)
23. Biggio, B., Fumera, G., Pillai, I., Roli, F.: A survey and experimental evaluation of image spam filtering techniques. Pattern Recogn. Lett. **32**, 1436–1446 (2011)
24. Fumera, G., Pillai, I., Roli, F.: Spam filtering based on the analysis of text information embedded into images. J. Mach. Learn. Res. **7**, 2699–2720 (2006)
25. Biggio, B., Fumera, G., Pillai, I., Roli, F.: Image spam filtering using visual information. In: Proceedings of the 14th International Conference on Image Analysis and Processing (ICIAP), pp. 105–110 (2007)
26. Mehta, B., Nangia, S., Gupta, M., Nejdl, W.: Detecting image spam using visual features and near duplicate detection. In: Proceedings of the 17th International Conference on World Wide Web, pp. 497–506 (2008)
27. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**(1/2), 177–196 (2001)
28. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, 8th European Conference on Computer Vision (ECCV), pp. 1–22 (2004)

29. Dredze, M., Gevaryahu, R., Elias-Bachrach, A.: Learning fast classifiers for image spam. In: Proceedings of the Conference on Email and Anti-Spam (CEAS), pp. 487–493 (2007)
30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
31. Bdiri, T., Bouguila, N.: An infinite mixture of inverted Dirichlet distributions. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011. LNCS, vol. 7063, pp. 71–78. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24958-7_9
32. Fan, W., Bouguila, N.: Topic novelty detection using infinite variational inverted Dirichlet mixture models. In: 14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, 9–11 December 2015, pp. 70–75 (2015)