

Variational Learning of a Dirichlet Process of Generalized Dirichlet Distributions for Simultaneous Clustering and Feature Selection

Wentao Fan^a, Nizar Bouguila^{b,*}

^a*Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada, H3G 1T7*

^b*The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada, H3G 1T7*

Abstract

This paper introduces a novel enhancement for unsupervised feature selection based on generalized Dirichlet (GD) mixture models. Our proposal is based on the extension of the finite mixture model previously developed in [1] to the infinite case, via the consideration of Dirichlet process mixtures, which can be viewed actually as a purely nonparametric model since the number of mixture components can increase as data are introduced. The infinite assumption is used to avoid problems related to model selection (i.e. determination of the number of clusters) and allows simultaneous separation of data in to similar clusters and selection of relevant features. Our resulting model is learned within a principled variational Bayesian framework that we have developed. The experimental results reported for both synthetic data and real-world challenging applications involving image categorization, automatic semantic annotation and retrieval show the ability of our approach to provide accurate models by distinguishing between relevant and irrelevant features without over- or under-fitting the data.

Keywords: Infinite mixture models, Dirichlet process, generalized Dirichlet, feature selection, clustering, images categorization, image auto-annotation.

1. Introduction

As the amount of multimedia information available increases, powerful approaches for analyzing, managing and categorizing these data become crucial. Clustering plays an important role in exploratory analysis of data. It provides principled means of discovering heterogenous groupings (i.e. clusters) in data and has been the topic of extensive research in the past [2, 3, 4, 5, 6, 7]. Data clustering is known to be a challenging task in modern knowledge discovery and data mining. This is especially true in high-dimensional spaces mainly because of data sparsity [8, 9] and a crucial step in this case is the selection of relevant features [10, 11, 12, 1]. Finite mixture models are well suited for clustering due to their simple structure and flexibility which offer a principled formal approach to unsupervised learning [13, 14]. In the classic approach to mixture models implementation, the density components are usually chosen as Gaussian and the number of components is supposed to be finite. Many methods for selecting the optimal number of clusters can be found in the literature (see, for instance, [15, 16, 17]). These approaches can be classified into two groups namely deterministic and Bayesian. The majority of both deterministic and Bayesian previous model selection approaches have to consider all possible values of the number of mixture components up to a certain maximum value and then choose the optimal one according to a certain criterion which is unfortunately computationally prohibitive (i.e. the learning algorithm have to be run for different choices of the number of mixture components) and may cause over- and under-fitting problems. A significant contribution that overcomes these drawbacks was made in [18] through the development of infinite mixture models which constitute an interesting extension of the typical finite mixture models approach by allowing the number of mixture components to increase as new data arrive. Infinite mixture models are based on the notion of Dirichlet processes which is one of the most popular Bayesian nonparametric models and is defined as a distribution over distributions [19, 20, 21]. Thanks to the the recent development

*Corresponding author: Tel.: +1 5148482424; Fax: +1 5148483171.

Email addresses: wenta_fa@encs.concordia.ca (Wentao Fan), nizar.bouguila@concordia.ca (Nizar Bouguila)

of Markov Chain Monte Carlo (MCMC) techniques [22], infinite mixture models have been widely and successfully used in various applications (see, for instance, [23, 24, 25, 26, 27, 28]) by embodying the well-known Occam’s Razor principle [29]. Concerning feature selection, although a lot of attention has been devoted to supervised feature selection (see, for instance, [30, 31, 32]), some unsupervised feature selection techniques have been proposed recently [33, 34, 35, 36, 37, 38, 39]. And some of these unsupervised techniques have been based on finite mixture models, but generally suppose that each per-component density is Gaussian with diagonal covariance matrix (i.e. the features are supposed independent)¹ [41, 42, 43].

Recently a nonparametric Bayesian unsupervised feature selection approach has been proposed in [44]. The main idea was the consideration of the infinite generalized Dirichlet (GD) mixture model, which offers high flexibility and ease of use, for simultaneous clustering and feature selection. One of the main advantages of this approach is that the structural properties of the GD allows it to be defined in a space where the independence of the features becomes a fact and not an assumption as shown for instance in [1]. The authors in [44] have proposed a fully Bayesian treatment of the unsupervised feature selection approach that they have previously introduced in [1] in order to overcome problems related to deterministic learning. The learning approach in [44] was based on the introduction of prior distributions over the mixture parameters. These parameters have been then estimated using a typical MCMC approach based on both Gibbs sampling and Metropolis-Hastings algorithms. MCMC techniques are effective for parameters estimation, but are unfortunately computationally very demanding and it can be very hard to diagnose their convergence. This is especially true in the case of high-dimensional data which involve the integration over a large number of model parameters. The accurate evaluation of such high-dimensional integrals has been the topic of extensive research. Recently, variational approaches, known also as ensemble learning [45, 46, 47], have been proposed as an efficient alternative to MCMC techniques. Motivated by the good results obtained recently using variational techniques for modeling mixture models, in this article we extend the learning approach in [44] by developing a variational alternative. The contribution of this paper is three-fold. First, we extend the finite GD mixture model with feature selection to the infinite case using a stick-breaking construction [48] such that the difficulty of choosing the appropriate number of clusters can be solved elegantly. Second, we propose a variational inference framework for learning the proposed model, such that the model parameters and features saliencies are estimated simultaneously in a closed form. In particular, conjugate priors are developed for all the involved parameters. Last, we apply the proposed approach to solve two challenging problems involving visual scenes categorization, and image automatic semantic annotation and retrieval. An appealing feature of the proposed variational approach is that it allows avoiding over-fitting by finding a compromise between generality and the number of parameters by implicitly providing a model order selection criterion [49, 46, 50]. Readers unfamiliar with Bayesian learning and the variational Bayes framework are referred to [45, 51].

The paper is organized as follows. In Section 2 we present our infinite feature selection model. In Section 3 we develop a practical variational approach to learn the parameters of this model. Section 4 is devoted to experimental results of using our approach. This is followed, in Section 5, by a discussion of our findings and conclusions.

2. The Infinite GD Mixture Model for Feature Selection

In this section, we describe our main unsupervised infinite feature selection model. We start by a brief overview of the finite GD mixture model. Then, the extension of this model to the infinite case and the integration of feature selection are proposed. Finally, we present the conjugate priors that we will consider for the resulting model learning.

2.1. The Finite GD Mixture Model

Consider a random vector $\vec{Y} = (Y_1, \dots, Y_D)$, drawn from a finite mixture of GD Distributions with M components [52] as

$$p(\vec{Y}|\vec{\pi}, \vec{\alpha}, \vec{\beta}) = \sum_{j=1}^M \pi_j \text{GD}(\vec{Y}|\vec{\alpha}_j, \vec{\beta}_j) \quad (1)$$

¹However, it is well-known that the independence assumption is infrequently met in practice [40].

where $\vec{\alpha} = \{\vec{\alpha}_1, \dots, \vec{\alpha}_M\}$, $\vec{\beta} = \{\vec{\beta}_1, \dots, \vec{\beta}_M\}$, $\vec{\alpha}_j$ and $\vec{\beta}_j$ are the parameters of the GD distribution representing component j with $\vec{\alpha}_j = \{\alpha_{j1}, \dots, \alpha_{jD}\}$ and $\vec{\beta}_j = \{\beta_{j1}, \dots, \beta_{jD}\}$, and $\vec{\pi} = \{\pi_1, \dots, \pi_M\}$ represents the mixing coefficients which are positive and sum to one. A GD distribution is defined as

$$\text{GD}(\vec{Y}|\vec{\alpha}_j, \vec{\beta}_j) = \prod_{l=1}^D \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} Y_l^{\alpha_{jl}-1} \left(1 - \sum_{k=1}^l Y_k\right)^{\beta_{jl}} \quad (2)$$

where $\sum_{l=1}^D Y_l < 1$ and $0 < Y_l < 1$ for $l = 1, \dots, D$, $\alpha_{jl} > 0$, $\beta_{jl} > 0$, $\gamma_{jl} = \beta_{jl} - \alpha_{j,l+1} - \beta_{j,l+1}$ for $l = 1, \dots, D-1$, and $\gamma_{jD} = \beta_{jD} - 1$.

Now, let us consider a set of N independent identically distributed vectors $\mathcal{Y} = (\vec{Y}_1, \dots, \vec{Y}_N)$ assumed to arise from a finite GD mixture. Following the Bayes' theorem, the probability that vector i is in cluster j conditional on having observed \vec{Y}_i (also known as *responsibilities*) can be written as

$$p(j|\vec{Y}_i) \propto \pi_j \text{GD}(\vec{Y}_i|\vec{\alpha}_j, \vec{\beta}_j) \quad (3)$$

In our work, we exploit an interesting mathematical property of the GD distribution previously discussed in [52, 1] to redefine the responsibilities as

$$p(j|\vec{Y}_i) \propto \pi_j \prod_{l=1}^D \text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl}) \quad (4)$$

where $X_{i1} = Y_{i1}$ and $X_{il} = Y_{il}/(1 - \sum_{k=1}^{l-1} Y_{ik})$ for $l > 1$ and $\text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl})$ is a Beta distribution defined with parameters $(\alpha_{jl}, \beta_{jl})$. Thus, the clustering structure for a finite GD mixture model underlying data set \mathcal{Y} can be represented by a new data set $\mathcal{X} = (\vec{X}_1, \dots, \vec{X}_N)$ using the following mixture model with conditionally independent features

$$p(\vec{X}_i|\vec{\pi}, \vec{\alpha}, \vec{\beta}) = \sum_{j=1}^M \pi_j \prod_{l=1}^D \text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl}) \quad (5)$$

It is noteworthy that this property plays a critical role for the GD mixture model, since the independence between the features becomes a fact rather than an assumption as considered in previous unsupervised feature selection Gaussian mixture-based approaches [41, 42].

2.2. Infinite GD Mixture Model With Feature Selection

The Dirichlet process (DP) [20] is a stochastic process whose sample paths are probability measures with probability one. It can be considered as a distribution over distributions. The infinite GD mixture model with feature selection proposed in this paper is constructed using the DP with a stick-breaking representation. Stick-breaking representation is an intuitive and straightforward constructive definition of the DP [48, 53, 54]. It is defined as follows: given a random distribution G , it is distributed according to a DP: $G \sim \text{DP}(\psi, H)$ if the following conditions are satisfied:

$$\lambda_j \sim \text{Beta}(1, \psi), \quad \Omega_j \sim H, \quad \pi_j = \lambda_j \prod_{s=1}^{j-1} (1 - \lambda_s), \quad G = \sum_{j=1}^{\infty} \pi_j \delta_{\Omega_j} \quad (6)$$

where δ_{Ω_j} denotes the Dirac delta measure centered at Ω_j , and ψ is a positive real number. The mixing weights π_j are obtained by recursively breaking an unit length stick into an infinite number of pieces.

Assuming now that the observed data set is generated from a GD mixture model with a countably infinite number of components. Thus, (5) can be rewritten as

$$p(\vec{X}_i|\vec{\pi}, \vec{\alpha}, \vec{\beta}) = \sum_{j=1}^{\infty} \pi_j \prod_{l=1}^D \text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl}). \quad (7)$$

Then, for each vector \vec{X}_i , we introduce a binary latent variable $\vec{Z}_i = (Z_{i1}, Z_{i2}, \dots)$, such $Z_{ij} \in \{0, 1\}$ and $Z_{ij} = 1$ if \vec{X}_i belongs to component j and 0, otherwise. Therefore, the likelihood function of the infinite GD mixture with latent

variables, which is actually the conditional distribution of data set \mathcal{X} given the class labels $\mathcal{Z} = (\vec{Z}_1, \dots, \vec{Z}_N)$ can be written as

$$p(\mathcal{X}|\mathcal{Z}, \vec{\alpha}, \vec{\beta}) = \prod_{i=1}^N \prod_{j=1}^{\infty} \left(\prod_{l=1}^D \text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl}) \right)^{Z_{ij}} \quad (8)$$

It is worth mentioning that all the features $\{X_{il}\}$ in the previous model are assumed to be equally important for the task of clustering which is not realistic in practice since some of the features might be irrelevant and do not contribute to the clustering process [1]. In order to take this fact into account the authors in [41] have supposed that a given feature X_{il} is generated from a mixture of two univariate distributions: The first one is assumed to generate relevant features and is different for each cluster; the second one is common to all clusters (i.e. independent from class labels) and assumed to generate irrelevant features. This idea has been extended in [1] where the irrelevant features are modeled as a finite mixture of distributions rather than a usual single distribution. In this work, we go a step further by modeling the irrelevant features with an infinite mixture model in order to bypass the difficulty of estimating the appropriate number of components for the mixture model representing irrelevant features. Therefore, each feature X_{il} can be approximated as

$$p(X_{il}) \simeq \left(\text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl}) \right)^{\phi_{il}} \left(\prod_{k=1}^{\infty} \text{Beta}(X_{il}|\sigma_{kl}, \tau_{kl})^{W_{ikl}} \right)^{1-\phi_{il}} \quad (9)$$

where W_{ikl} is a binary variable such that $W_{ikl} = 1$ if X_{il} comes from the k th component of the infinite Beta mixture for the irrelevant features. ϕ_{il} is a binary latent variable, such that $\phi_{il} = 1$ indicates that feature l is relevant and follows a Beta distribution $\text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl})$, and $\phi_{il} = 0$ denotes that feature l is irrelevant and supposed to follow an infinite mixture of Beta distributions independent from the class labels:

$$p(X_{il}) = \sum_{k=1}^{\infty} \eta_k \text{Beta}(X_{il}|\sigma_{kl}, \tau_{kl}) \quad (10)$$

where η_k denotes the mixing probability and also implies the prior probability that X_{il} is generated from the k th component of the infinite Beta mixture representing irrelevant features.

Thus, we can write the likelihood of the observed data set \mathcal{X} following the infinite GD mixture model with feature selection as

$$p(\mathcal{X}|\mathcal{Z}, \mathcal{W}, \vec{\phi}, \vec{\alpha}, \vec{\beta}, \vec{\sigma}, \vec{\tau}) = \prod_{i=1}^N \prod_{j=1}^{\infty} \left[\prod_{l=1}^D \text{Beta}(X_{il}|\alpha_{jl}, \beta_{jl})^{\phi_{il}} \times \left(\prod_{k=1}^{\infty} \text{Beta}(X_{il}|\sigma_{kl}, \tau_{kl})^{W_{ikl}} \right)^{1-\phi_{il}} \right]^{Z_{ij}} \quad (11)$$

where $\mathcal{W} = (\vec{W}_1, \dots, \vec{W}_N)$ with $\vec{W}_i = (\vec{W}_{i1}, \vec{W}_{i2}, \dots)$ and $\vec{W}_{ik} = (W_{ik1}, \dots, W_{ikD})$. $\vec{\phi} = (\vec{\phi}_1, \dots, \vec{\phi}_N)$ contains elements $\vec{\phi}_i = (\phi_{i1}, \dots, \phi_{iD})$. $\vec{\sigma} = (\vec{\sigma}_1, \vec{\sigma}_2, \dots)$ and $\vec{\tau} = (\vec{\tau}_1, \vec{\tau}_2, \dots)$ are the parameters of the Beta mixture representing irrelevant features which comprise elements $\vec{\sigma}_k = (\sigma_{k1}, \dots, \sigma_{kD})$ and $\vec{\tau}_k = (\tau_{k1}, \dots, \tau_{kD})$, respectively. The main idea of the unsupervised feature selection method in our work is shown in Figure 1. The merits of adopting this feature selection technique shall be demonstrated through experiments in Section 4. For more details about this unsupervised feature selection model, the reader is referred to [41, 1].

2.3. Prior Distributions of The Proposed Model

We shall follow a variational Bayesian approach for learning our model, thus each unknown parameter is given a prior distribution. In our work, we choose conjugate priors for the unknown random variables $\mathcal{Z}, \mathcal{W}, \vec{\phi}, \vec{\alpha}, \vec{\beta}, \vec{\sigma}$ and $\vec{\tau}$. The consideration of conjugate prior distributions is motivated by the fact that it may lead to a considerably simplified Bayesian analysis in which the posterior distributions have the same functional forms as the priors. More importantly, the whole variational inference process becomes tractable and closed form solutions for updating optimal factors can be obtained when using conjugate priors in conjunction with the factorization assumption, as we shall see in the next section. The prior distributions of \mathcal{Z} and \mathcal{W} given the mixing coefficients $\vec{\pi}$ and $\vec{\eta}$ can be specified as

$$p(\mathcal{Z}|\vec{\pi}) = \prod_{i=1}^N \prod_{j=1}^{\infty} \pi_j^{Z_{ij}} \quad (12)$$

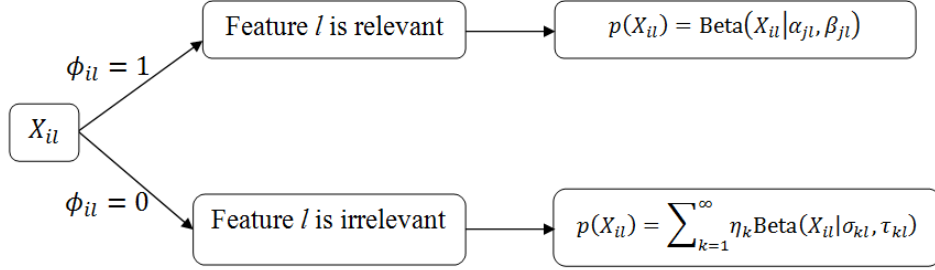


Figure 1: The unsupervised feature selection method: when $\phi_{il} = 1$, feature X_{il} is relevant and follows a Beta distribution $\text{Beta}(X_{il} | \alpha_{jl}, \beta_{jl})$; when $\phi_{il} = 0$, feature X_{il} is irrelevant and follows an infinite mixture of Beta distributions: $\sum_{k=1}^{\infty} \eta_k \text{Beta}(X_{il} | \sigma_{kl}, \tau_{kl})$.

$$p(\mathcal{W} | \vec{\eta}) = \prod_{i=1}^N \prod_{k=1}^{\infty} \prod_{l=1}^D \eta_k^{W_{ikl}} \quad (13)$$

According to the stick-breaking construction of DP as stated in (6), $\vec{\pi}$ is a function of $\vec{\lambda}$. We rewrite it here for the sake of clarity

$$\pi_j = \lambda_j \prod_{s=1}^{j-1} (1 - \lambda_s) \quad (14)$$

Similarly, $\vec{\eta}$ can be defined as a function of $\vec{\gamma}$, such that

$$\eta_k = \gamma_k \prod_{s=1}^{k-1} (1 - \gamma_s) \quad (15)$$

Therefore, we can rewrite (12) and (13) as

$$p(\mathcal{Z} | \vec{\lambda}) = \prod_{i=1}^N \prod_{j=1}^{\infty} [\lambda_j \prod_{s=1}^{j-1} (1 - \lambda_s)]^{Z_{ij}} \quad p(\mathcal{W} | \vec{\gamma}) = \prod_{i=1}^N \prod_{k=1}^{\infty} \prod_{l=1}^D [\gamma_k \prod_{s=1}^{k-1} (1 - \gamma_s)]^{W_{ikl}} \quad (16)$$

where $\vec{\lambda} = (\lambda_1, \lambda_2, \dots)$ and $\vec{\gamma} = (\gamma_1, \gamma_2, \dots)$. The prior distributions of $\vec{\lambda}$ and $\vec{\gamma}$ follow the specific Beta distribution given in (6) as

$$p(\vec{\lambda} | \vec{\psi}) = \prod_{j=1}^{\infty} \text{Beta}(1, \psi_j) = \prod_{j=1}^{\infty} \psi_j (1 - \lambda_j)^{\psi_j - 1} \quad (17)$$

$$p(\vec{\gamma} | \vec{\varphi}) = \prod_{k=1}^{\infty} \text{Beta}(1, \varphi_k) = \prod_{k=1}^{\infty} \varphi_k (1 - \gamma_k)^{\varphi_k - 1} \quad (18)$$

To add more flexibility, another layer is added to the Bayesian hierarchy by introducing prior distributions over the hyperparameters $\vec{\psi} = (\psi_1, \psi_2, \dots)$ and $\vec{\varphi} = (\varphi_1, \varphi_2, \dots)$. Motivated by the fact that the Gamma distribution is conjugate to the stick lengths [47], Gamma priors are placed over $\vec{\psi}$ and $\vec{\varphi}$ as

$$p(\vec{\psi}) = \mathcal{G}(\vec{\psi} | \vec{a}, \vec{b}) = \prod_{j=1}^{\infty} \frac{b_j^{a_j}}{\Gamma(a_j)} \psi_j^{a_j - 1} e^{-b_j \psi_j} \quad p(\vec{\varphi}) = \mathcal{G}(\vec{\varphi} | \vec{c}, \vec{d}) = \prod_{k=1}^{\infty} \frac{d_k^{c_k}}{\Gamma(c_k)} \varphi_k^{c_k - 1} e^{-d_k \varphi_k} \quad (19)$$

where hyperparameters $\vec{a} = (a_1, a_2, \dots)$, $\vec{b} = (b_1, b_2, \dots)$, $\vec{c} = (c_1, c_2, \dots)$ and $\vec{d} = (d_1, d_2, \dots)$ are subject to the constraints $a_j > 0$, $b_j > 0$, $c_k > 0$ and $d_k > 0$ to ensure that these two prior distributions can be normalized. The prior distribution for the feature relevance indicator variable $\vec{\phi}$ is defined as

$$p(\vec{\phi} | \vec{\epsilon}) = \prod_{i=1}^N \prod_{l=1}^D \epsilon_{l1}^{\phi_{il}} \epsilon_{l2}^{1 - \phi_{il}} \quad (20)$$

where each ϕ_{il} is a Bernoulli variable such that $p(\phi_{il} = 1) = \epsilon_{l_1}$ and $p(\phi_{il} = 0) = \epsilon_{l_2}$. The vector $\vec{\epsilon} = (\vec{\epsilon}_1, \dots, \vec{\epsilon}_D)$ represents the features saliencies (i.e. the probabilities that the features are relevant) such that $\vec{\epsilon}_l = (\epsilon_{l_1}, \epsilon_{l_2})$ and $\epsilon_{l_1} + \epsilon_{l_2} = 1$. Furthermore, a Dirichlet distribution is chosen over $\vec{\epsilon}$ as [55]

$$p(\vec{\epsilon}) = \prod_{l=1}^D \text{Dir}(\vec{\epsilon}_l | \vec{\xi}) = \prod_{l=1}^D \frac{\Gamma(\xi_1 + \xi_2)}{\Gamma(\xi_1)\Gamma(\xi_2)} \epsilon_{l_1}^{\xi_1-1} \epsilon_{l_2}^{\xi_2-1} \quad (21)$$

where the hyperparameter $\vec{\xi} = (\xi_1, \xi_2)$ is subject to the constraint $(\xi_1, \xi_2) > 0$ in order to ensure that the distribution can be normalized. Next, we need to define the prior distributions for parameters $\vec{\alpha}$, $\vec{\beta}$, $\vec{\sigma}$ and $\vec{\tau}$ of Beta distributions. Although Beta distribution belongs to the exponential family and has a formal conjugate prior [46], it is analytically intractable and cannot be used within a variational framework as shown for instance in [56]. Thus, the Gamma distribution is adopted to approximate the conjugate prior, as suggested in [56], by assuming that parameters of Beta distributions are statistically independent:

$$p(\vec{\alpha}) = \mathcal{G}(\vec{\alpha} | \vec{u}, \vec{v}) = \prod_{j=1}^{\infty} \prod_{l=1}^D \frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-v_{jl}\alpha_{jl}} \quad (22)$$

$$p(\vec{\beta}) = \mathcal{G}(\vec{\beta} | \vec{p}, \vec{q}) = \prod_{j=1}^{\infty} \prod_{l=1}^D \frac{q_{jl}^{p_{jl}}}{\Gamma(p_{jl})} \beta_{jl}^{p_{jl}-1} e^{-q_{jl}\beta_{jl}} \quad (23)$$

$$p(\vec{\sigma}) = \mathcal{G}(\vec{\sigma} | \vec{g}, \vec{h}) = \prod_{k=1}^{\infty} \prod_{l=1}^D \frac{h_{kl}^{g_{kl}}}{\Gamma(g_{kl})} \sigma_{kl}^{g_{kl}-1} e^{-h_{kl}\sigma_{kl}} \quad (24)$$

$$p(\vec{\tau}) = \mathcal{G}(\vec{\tau} | \vec{s}, \vec{t}) = \prod_{k=1}^{\infty} \prod_{l=1}^D \frac{t_{kl}^{s_{kl}}}{\Gamma(s_{kl})} \tau_{kl}^{s_{kl}-1} e^{-t_{kl}\tau_{kl}} \quad (25)$$

where all the hyperparameters $\vec{u} = \{u_{jl}\}$, $\vec{v} = \{v_{jl}\}$, $\vec{p} = \{p_{jl}\}$, $\vec{q} = \{q_{jl}\}$, $\vec{g} = \{g_{kl}\}$, $\vec{h} = \{h_{kl}\}$, $\vec{s} = \{s_{kl}\}$ and $\vec{t} = \{t_{kl}\}$ of the above conjugate priors are positive. In our work, we define $\Theta = \{\mathcal{Z}, \mathcal{W}, \vec{\phi}, \vec{\alpha}, \vec{\beta}, \vec{\sigma}, \vec{\tau}, \vec{\lambda}, \vec{\psi}, \vec{\gamma}, \vec{\phi}, \vec{\epsilon}\}$ as the set of unknown random variables. After defining the priors for all the unknown variables in the proposed model, the joint distribution of all the random variables is given by

$$\begin{aligned} p(\mathcal{X}, \Theta) &= p(\mathcal{X} | \mathcal{Z}, \mathcal{W}, \vec{\phi}, \vec{\alpha}, \vec{\beta}, \vec{\sigma}, \vec{\tau}) p(\mathcal{Z} | \vec{\lambda}) p(\vec{\lambda} | \vec{\psi}) p(\vec{\psi}) p(\mathcal{W} | \vec{\gamma}) p(\vec{\gamma} | \vec{\phi}) p(\vec{\phi}) p(\vec{\epsilon}) p(\vec{\alpha}) p(\vec{\beta}) p(\vec{\sigma}) p(\vec{\tau}) \\ &= \prod_{i=1}^N \prod_{j=1}^{\infty} \left\{ \prod_{l=1}^D \left[\frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} X_{il}^{\alpha_{jl}-1} (1 - X_{il})^{\beta_{jl}-1} \right]^{\phi_{il}} \left[\prod_{k=1}^{\infty} \left(\frac{\Gamma(\sigma_{kl} + \tau_{kl})}{\Gamma(\sigma_{kl})\Gamma(\tau_{kl})} X_{il}^{\sigma_{kl}-1} (1 - X_{il})^{\tau_{kl}-1} \right)^{W_{ikl}} \right]^{1-\phi_{il}} \right\}^{Z_{ij}} \\ &\times \prod_{i=1}^N \prod_{j=1}^{\infty} \prod_{s=1}^{j-1} [\lambda_j (1 - \lambda_s)]^{Z_{ij}} \times \prod_{j=1}^{\infty} \psi_j (1 - \lambda_j)^{\psi_j-1} \prod_{j=1}^{\infty} \frac{b_j^{a_j}}{\Gamma(a_j)} \psi_j^{a_j-1} e^{-b_j \psi_j} \times \prod_{i=1}^N \prod_{k=1}^{\infty} \prod_{l=1}^D [\gamma_k \prod_{s=1}^{k-1} (1 - \gamma_s)]^{W_{ikl}} \\ &\times \prod_{k=1}^{\infty} \varphi_k (1 - \gamma_k)^{\varphi_k-1} \times \prod_{k=1}^{\infty} \frac{d_k^{c_k}}{\Gamma(c_k)} \varphi_k^{c_k-1} e^{-d_k \varphi_k} \prod_{i=1}^N \prod_{l=1}^D \epsilon_{l_1}^{\phi_{il}} \epsilon_{l_2}^{1-\phi_{il}} \times \prod_{l=1}^D \frac{\Gamma(\xi_1 + \xi_2)}{\Gamma(\xi_1)\Gamma(\xi_2)} \epsilon_{l_1}^{\xi_1-1} \epsilon_{l_2}^{\xi_2-1} \\ &\times \prod_{j=1}^{\infty} \prod_{l=1}^D \left[\frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} e^{-v_{jl}\alpha_{jl}} \frac{q_{jl}^{p_{jl}}}{\Gamma(p_{jl})} \beta_{jl}^{p_{jl}-1} e^{-q_{jl}\beta_{jl}} \right] \prod_{k=1}^{\infty} \prod_{l=1}^D \left[\frac{h_{kl}^{g_{kl}}}{\Gamma(g_{kl})} \sigma_{kl}^{g_{kl}-1} e^{-h_{kl}\sigma_{kl}} \frac{t_{kl}^{s_{kl}}}{\Gamma(s_{kl})} \tau_{kl}^{s_{kl}-1} e^{-t_{kl}\tau_{kl}} \right] \end{aligned} \quad (26)$$

A directed graphical representation of this model is illustrated in Figure 2.

3. Variational Inference

In this section, a variational framework for learning the infinite GD mixture model with feature selection is proposed. The main idea in variational learning is to find an approximation for the posterior distribution $p(\Theta | \mathcal{X})$ as well as for the model evidence $p(\mathcal{X})$ [50]. First, the log marginal probability $\ln p(\mathcal{X})$ can be decomposed as

$$\ln p(\mathcal{X}) = \underbrace{\int Q(\Theta) \ln \frac{p(\mathcal{X}, \Theta)}{Q(\Theta)} d\Theta}_{\mathcal{L}(Q)} - \underbrace{\int Q(\Theta) \ln \frac{p(\Theta | \mathcal{X})}{Q(\Theta)} d\Theta}_{\text{KL}(Q(\Theta) \| p(\Theta | \mathcal{X}))} \quad (27)$$

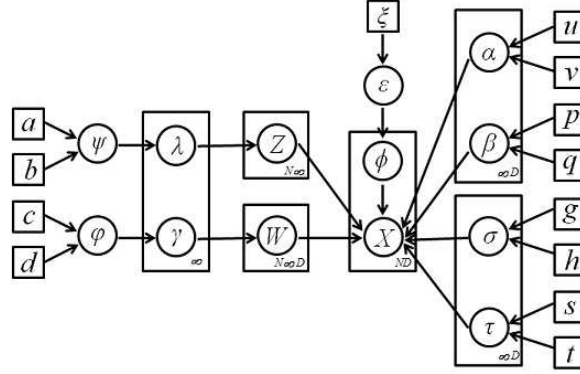


Figure 2: Graphical model representation of the infinite GD mixture model with feature selection. Symbols in circles denote random variables; otherwise, they denote model parameters. Plates indicate repetition (with the number of repetitions in the lower right), and arcs describe conditional dependencies between variables.

where $Q(\Theta)$ is an approximation to the true posterior distribution $p(\Theta|\mathcal{X})$. The second term on the right hand side of (27) is the Kullback-Leibler (KL) divergence between $Q(\Theta)$ and the posterior distribution $p(\Theta|\mathcal{X})$. In order to verify the decomposition in (27), we can use the fact that $\ln p(\mathcal{X}, \Theta) = \ln p(\Theta|\mathcal{X}) + \ln p(\mathcal{X})$. After we substitute this decomposition of $\ln p(\mathcal{X}, \Theta)$ into the expression of $\mathcal{L}(Q)$ in (27), then we can have two terms: one of which cancels $\text{KL}(Q(\Theta) \| p(\Theta|\mathcal{X}))$ while the other results in the required log likelihood $\ln p(\mathcal{X})$. Since $\text{KL}(Q(\Theta) \| p(\Theta|\mathcal{X})) \geq 0$ (with equality if, and only if $Q(\Theta) = p(\Theta|\mathcal{X})$), it is obvious that $\mathcal{L}(Q) \leq \ln p(\mathcal{X})$. Therefore, $\mathcal{L}(Q)$ can be considered as a lower bound for $\ln p(\mathcal{X})$ [57]. Obviously, this lower bound is maximized when the KL divergence vanishes, that is when $Q(\Theta)$ equals the true posterior distribution $p(\Theta|\mathcal{X})$. Nevertheless, in practice the true posterior distribution is normally computationally intractable and cannot be directly adopted in variational inference. Thus, in this work, we exploit a factorization assumption which is known as *mean field theory* for restricting the form of $Q(\Theta)$. This approximation framework has been used efficiently for variational inference by several researchers in the past [51, 50]. Under this assumption, the posterior distribution $Q(\Theta)$ can be factorized into disjoint tractable distributions such that $Q(\Theta) = \prod_i Q_i(\Theta_i)$. It is worth mentioning that this assumption is imposed purely to achieve tractability. Moreover, this is the only assumption about the distribution, and no restriction is placed on the functional forms of the individual factors $Q_i(\Theta_i)$. To maximize the lower bound $\mathcal{L}(Q)$, we need to make a variational optimization of $\mathcal{L}(Q)$ with respect to each of the factor distributions $Q_i(\Theta_i)$ in turn. Indeed, for a specific factor $Q_s(\Theta_s)$ in a standard variational inference framework, the general expression for its optimal solution is given by [51, 58, 50]

$$Q_s(\Theta_s) = \frac{\exp(\langle \ln p(\mathcal{X}, \Theta) \rangle_{i \neq s})}{\int \exp(\langle \ln p(\mathcal{X}, \Theta) \rangle_{i \neq s}) d\Theta} \quad (28)$$

where $\langle \cdot \rangle_{i \neq s}$ denotes an expectation with respect to all the distributions $Q_i(\Theta_i)$ except for $i = s$. In variational inference, all factors $Q_i(\Theta_i)$ need to be suitably initialized first, then each individual factor is updated in turn with a revised value obtained by (28) using the current values of all the other factors. Furthermore, we truncate the stick-breaking representation for the infinite GD mixture model at a value of M as

$$\lambda_M = 1, \quad \pi_j = 0 \text{ when } j > M, \quad \sum_{j=1}^M \pi_j = 1 \quad (29)$$

Moreover, the infinite Beta mixture model for the irrelevant features is truncated at a value of K such that

$$\gamma_K = 1, \quad \eta_k = 0 \text{ when } k > K, \quad \sum_{k=1}^K \eta_k = 1 \quad (30)$$

Note that, the truncation levels M and K are variational parameters which can be freely initialized and will be optimized automatically during the learning process. By employing the factorization assumption and the truncated

stick-breaking representation for the proposed model, we then obtain

$$\begin{aligned} Q(\Theta) = & \left[\prod_{i=1}^N \prod_{j=1}^M Q(Z_{ij}) \right] \left[\prod_{j=1}^M Q(\lambda_j) Q(\psi_j) \right] \left[\prod_{i=1}^N \prod_{k=1}^K \prod_{l=1}^D Q(W_{ikl}) \right] \left[\prod_{k=1}^K Q(\gamma_k) Q(\varphi_k) \right] \left[\prod_{i=1}^N \prod_{l=1}^D Q(\phi_{il}) \right] \left[\prod_{l=1}^D Q(\vec{\epsilon}_l) \right] \\ & \times \left[\prod_{j=1}^M \prod_{l=1}^D Q(\alpha_{jl}) Q(\beta_{jl}) \right] \left[\prod_{k=1}^K \prod_{l=1}^D Q(\sigma_{kl}) Q(\tau_{kl}) \right] \end{aligned} \quad (31)$$

By applying (28) to each factor of the variational posterior, we then acquire the following optimal solutions (details of the variational inference are given in Appendix I)

$$Q(\mathcal{Z}) = \prod_{i=1}^N \prod_{j=1}^M r_{ij}^{Z_{ij}}, \quad Q(\vec{\lambda}) = \prod_{j=1}^M \text{Beta}(\lambda_j | \theta_j, \vartheta_j) \quad (32)$$

$$Q(\vec{\psi}) = \prod_{j=1}^M \mathcal{G}(\psi_j | a_j^*, b_j^*), \quad Q(\mathcal{W}) = \prod_{i=1}^N \prod_{k=1}^K \prod_{l=1}^D (m_{ikl}^{W_{ikl}}) \quad (33)$$

$$Q(\vec{\gamma}) = \prod_{k=1}^K \text{Beta}(\gamma_k | \rho_k, \varpi_k), \quad Q(\vec{\varphi}) = \prod_{k=1}^K \mathcal{G}(\varphi_k | c_k^*, d_k^*) \quad (34)$$

$$Q(\vec{\phi}) = \prod_{i=1}^N \prod_{l=1}^D f_{il}^{\phi_{il}} (1 - f_{il})^{(1-\phi_{il})}, \quad Q(\vec{\epsilon}) = \prod_{l=1}^D \text{Dir}(\vec{\epsilon}_l | \xi^*) \quad (35)$$

$$Q(\vec{\alpha}) = \prod_{j=1}^M \prod_{l=1}^D \mathcal{G}(\alpha_{jl} | u_{jl}^*, v_{jl}^*), \quad Q(\vec{\beta}) = \prod_{j=1}^M \prod_{l=1}^D \mathcal{G}(\beta_{jl} | p_{jl}^*, q_{jl}^*) \quad (36)$$

$$Q(\vec{\sigma}) = \prod_{k=1}^K \prod_{l=1}^D \mathcal{G}(\sigma_{kl} | g_{kl}^*, h_{kl}^*), \quad Q(\vec{\tau}) = \prod_{k=1}^K \prod_{l=1}^D \mathcal{G}(\tau_{kl} | s_{kl}^*, t_{kl}^*) \quad (37)$$

where we have defined

$$r_{ij} = \frac{\tilde{r}_{ij}}{\sum_{j=1}^M \tilde{r}_{ij}}, \quad f_{il} = \frac{f_{il}^{\phi_{il}}}{f_{il}^{\phi_{il}} + f_{il}^{(1-\phi_{il})}}, \quad m_{ikl} = \frac{\tilde{m}_{ikl}}{\sum_{k=1}^K \tilde{m}_{ikl}} \quad (38)$$

$$\tilde{r}_{ij} = \exp \left\{ \sum_{l=1}^D \langle \phi_{il} \rangle [\tilde{\mathcal{R}}_{jl} + (\tilde{\alpha}_{jl} - 1) \ln X_{il} + (\tilde{\beta}_{jl} - 1) \ln(1 - X_{il})] + \langle \ln \lambda_j \rangle + \sum_{s=1}^{j-1} \langle \ln(1 - \lambda_s) \rangle \right\} \quad (39)$$

$$\tilde{m}_{ikl} = \exp \left\{ \langle 1 - \phi_{il} \rangle [\tilde{\mathcal{F}}_{kl} + (\tilde{\sigma}_{kl} - 1) \ln X_{il} + (\tilde{\tau}_{kl} - 1) \ln(1 - X_{il})] + \langle \ln \gamma_k \rangle + \sum_{s=1}^{k-1} \langle \ln(1 - \gamma_s) \rangle \right\} \quad (40)$$

$$f_{il}^{\phi_{il}} = \exp \left\{ \sum_{j=1}^M \langle Z_{ij} \rangle [\tilde{\mathcal{R}}_{jl} + (\tilde{\alpha}_{jl} - 1) \ln X_{il} + (\tilde{\beta}_{jl} - 1) \ln(1 - X_{il})] + \langle \ln \epsilon_{1i} \rangle \right\} \quad (41)$$

$$f_{il}^{(1-\phi_{il})} = \exp \left\{ \sum_{k=1}^K \langle W_{ikl} \rangle [\tilde{\mathcal{F}}_{kl} + (\tilde{\sigma}_{kl} - 1) \ln X_{il} + (\tilde{\tau}_{kl} - 1) \ln(1 - X_{il})] + \langle \ln \epsilon_{2i} \rangle \right\} \quad (42)$$

$$\begin{aligned} \tilde{\mathcal{R}} = & \ln \frac{\Gamma(\tilde{\alpha} + \tilde{\beta})}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})} + \tilde{\alpha} [\Psi(\tilde{\alpha} + \tilde{\beta}) - \Psi(\tilde{\alpha})] \langle \ln \alpha \rangle - \ln \tilde{\alpha} + \tilde{\beta} [\Psi(\tilde{\alpha} + \tilde{\beta}) - \Psi(\tilde{\beta})] \langle \ln \beta \rangle - \ln \tilde{\beta} \\ & + 0.5 \tilde{\alpha}^2 [\Psi'(\tilde{\alpha} + \tilde{\beta}) - \Psi'(\tilde{\alpha})] \langle (\ln \alpha - \ln \tilde{\alpha})^2 \rangle + 0.5 \tilde{\beta}^2 [\Psi'(\tilde{\alpha} + \tilde{\beta}) - \Psi'(\tilde{\beta})] \langle (\ln \beta - \ln \tilde{\beta})^2 \rangle \\ & + \tilde{\alpha} \tilde{\beta} \Psi'(\tilde{\alpha} + \tilde{\beta}) \langle \ln \alpha \rangle - \ln \tilde{\alpha} \langle \ln \beta \rangle - \ln \tilde{\beta} \end{aligned} \quad (43)$$

$$\begin{aligned} \tilde{\mathcal{F}} = & \ln \frac{\Gamma(\tilde{\sigma} + \tilde{\tau})}{\Gamma(\tilde{\sigma})\Gamma(\tilde{\tau})} + \tilde{\sigma} [\Psi(\tilde{\sigma} + \tilde{\tau}) - \Psi(\tilde{\sigma})] \langle \ln \sigma \rangle - \ln \tilde{\sigma} + \tilde{\tau} [\Psi(\tilde{\sigma} + \tilde{\tau}) - \Psi(\tilde{\tau})] \langle \ln \tau \rangle - \ln \tilde{\tau} \\ & + 0.5 \tilde{\sigma}^2 [\Psi'(\tilde{\sigma} + \tilde{\tau}) - \Psi'(\tilde{\sigma})] \langle (\ln \sigma - \ln \tilde{\sigma})^2 \rangle + 0.5 \tilde{\tau}^2 [\Psi'(\tilde{\sigma} + \tilde{\tau}) - \Psi'(\tilde{\tau})] \langle (\ln \tau - \ln \tilde{\tau})^2 \rangle \\ & + \tilde{\sigma} \tilde{\tau} \Psi'(\tilde{\sigma} + \tilde{\tau}) \langle \ln \sigma \rangle - \ln \tilde{\sigma} \langle \ln \tau \rangle - \ln \tilde{\tau} \end{aligned} \quad (44)$$

$$u_{jl}^* = u_{jl} + \sum_{i=1}^N \langle Z_{ij} \rangle \langle \phi_{il} \rangle \bar{\alpha}_{jl} [\Psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \Psi(\bar{\alpha}_{jl}) + \bar{\beta}_{jl} \Psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) (\langle \ln \beta_{jl} \rangle - \ln \bar{\beta}_{jl})] \quad (45)$$

$$p_{jl}^* = p_{jl} + \sum_{i=1}^N \langle Z_{ij} \rangle \langle \phi_{il} \rangle \bar{\beta}_{jl} [\Psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \Psi(\bar{\beta}_{jl}) + \bar{\alpha}_{jl} \Psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) (\langle \ln \alpha_{jl} \rangle - \ln \bar{\alpha}_{jl})] \quad (46)$$

$$g_{kl}^* = g_{kl} + \sum_{i=1}^N (1 - \phi_{il}) \langle W_{ikl} \rangle \bar{\sigma}_{kl} [\Psi(\bar{\sigma}_{kl} + \bar{\tau}_{kl}) - \Psi(\bar{\sigma}_{kl}) + \bar{\tau}_{kl} \Psi'(\bar{\sigma}_{kl} + \bar{\tau}_{kl}) (\langle \ln \tau_{kl} \rangle - \ln \bar{\tau}_{kl})] \quad (47)$$

$$s_{kl}^* = s_{kl} + \sum_{i=1}^N (1 - \phi_{il}) \langle W_{ikl} \rangle \bar{\tau}_{kl} [\Psi(\bar{\sigma}_{kl} + \bar{\tau}_{kl}) - \Psi(\bar{\tau}_{kl}) + \bar{\sigma}_{kl} \Psi'(\bar{\sigma}_{kl} + \bar{\tau}_{kl}) (\langle \ln \sigma_{kl} \rangle - \ln \bar{\sigma}_{kl})] \quad (48)$$

$$v_{jl}^* = v_{jl} - \sum_{i=1}^N \langle Z_{ij} \rangle \langle \phi_{il} \rangle \ln X_{il}, \quad q_{jl}^* = q_{jl} - \sum_{i=1}^N \langle Z_{ij} \rangle \langle \phi_{il} \rangle \ln(1 - X_{il}) \quad (49)$$

$$h_{kl}^* = h_{kl} - \sum_{i=1}^N (1 - \phi_{il}) \langle W_{ikl} \rangle \ln X_{il}, \quad t_{kl}^* = t_{kl} - \sum_{i=1}^N (1 - \phi_{il}) \langle W_{ikl} \rangle \ln(1 - X_{il}) \quad (50)$$

$$\theta_j = 1 + \sum_{i=1}^N \langle Z_{ij} \rangle, \quad \vartheta_j = \langle \psi \rangle + \sum_{i=1}^N \sum_{s=j+1}^M \langle Z_{is} \rangle, \quad a_j^* = a_j + 1 \quad (51)$$

$$b_j^* = b_j - \langle \ln(1 - \lambda_j) \rangle, \quad \rho_k = 1 + \sum_{i=1}^N \sum_{l=1}^D \langle W_{ikl} \rangle, \quad c_k^* = c_k + 1 \quad (52)$$

$$\varpi_k = \langle \varphi_k \rangle + \sum_{i=1}^N \sum_{s=k+1}^K \sum_{l=1}^D \langle W_{isl} \rangle, \quad d_k^* = d_k - \langle \ln(1 - \gamma_k) \rangle \quad (53)$$

$$\xi_1^* = \xi_1 + \sum_{i=1}^N \langle \phi_{il} \rangle, \quad \xi_2^* = \xi_2 + \sum_{i=1}^N (1 - \phi_{il}) \quad (54)$$

where $\Psi(\cdot)$ is the digamma function and defined as: $\Psi(a) = d \ln \Gamma(a) / da$. The expected values in the above formulas are given by

$$\bar{\alpha}_{jl} = \frac{u_{jl}^*}{v_{jl}^*}, \quad \bar{\beta}_{jl} = \frac{p_{jl}^*}{q_{jl}^*}, \quad \bar{\sigma}_{kl} = \frac{g_{kl}^*}{h_{kl}^*}, \quad \bar{\tau}_{kl} = \frac{s_{kl}^*}{t_{kl}^*} \quad (55)$$

$$\langle \psi_j \rangle = \frac{a_j^*}{\vartheta_j}, \quad \langle \varphi_k \rangle = \frac{c_k^*}{d_k^*}, \quad \langle Z_{ij} \rangle = r_{ij}, \quad \langle W_{ikl} \rangle = m_{ikl} \quad (56)$$

$$\langle \phi_{il} \rangle = f_{il}, \quad \langle 1 - \phi_{il} \rangle = 1 - f_{il}, \quad \langle \ln \alpha \rangle = \Psi(u^*) - \ln v^* \quad (57)$$

$$\langle \ln \beta \rangle = \Psi(p^*) - \ln q^*, \quad \langle \ln \sigma \rangle = \Psi(g^*) - \ln h^*, \quad \langle \ln \tau \rangle = \Psi(s^*) - \ln t^* \quad (58)$$

$$\langle \ln \lambda_j \rangle = \Psi(\theta) - \Psi(\theta + \vartheta), \quad \langle \ln(1 - \lambda_j) \rangle = \Psi(\vartheta) - \Psi(\theta + \vartheta) \quad (59)$$

$$\langle \ln \gamma_k \rangle = \Psi(\rho) - \Psi(\rho + \varpi), \quad \langle \ln(1 - \gamma_k) \rangle = \Psi(\varpi) - \Psi(\rho + \varpi) \quad (60)$$

$$\langle \ln \epsilon_1 \rangle = \Psi(\xi_1^*) - \Psi(\xi_1^* + \xi_2^*), \quad \langle \ln \epsilon_2 \rangle = \Psi(\xi_2^*) - \Psi(\xi_1^* + \xi_2^*) \quad (61)$$

$$\langle (\ln \alpha - \ln \bar{\alpha})^2 \rangle = [\Psi(u^*) - \ln u^*]^2 + \Psi'(u^*) \quad (62)$$

$$\langle (\ln \beta - \ln \bar{\beta})^2 \rangle = [\Psi(p^*) - \ln p^*]^2 + \Psi'(p^*) \quad (63)$$

$$\langle (\ln \sigma - \ln \bar{\sigma})^2 \rangle = [\Psi(g^*) - \ln g^*]^2 + \Psi'(g^*) \quad (64)$$

$$\langle (\ln \tau - \ln \bar{\tau})^2 \rangle = [\Psi(s^*) - \ln s^*]^2 + \Psi'(s^*) \quad (65)$$

Since the solutions to each variational factor are coupled together through the expected values of other factors, the optimization of the model can be solved in a way analogous to the EM algorithm. In the variational equivalent of the E-step, we optimize the moments using the current model parameters through (55)~(65). Then, in the subsequent variational equivalent of the M-step, we keep the values of those moments fixed and use them to re-estimate the variational distributions by (32)~(37). These two steps are repeated until convergence. The complete learning process is summarized in Algorithm 1.²

²The complete source code is available upon request.

Algorithm 1 Variational learning of infinite GD mixtures with feature selection

- 1: Choose the initial truncation levels M and K .
 - 2: Initialize the values for hyper-parameters $u_{jl}, v_{jl}, p_{jl}, q_{jl}, g_{kl}, h_{kl}, s_{kl}, t_{kl}, a_j, b_j, c_k, d_k, \xi_1$ and ξ_2 .
 - 3: Initialize the values of r_{ij} and m_{ikl} by K -Means algorithm.
 - 4: **repeat**
 - 5: The variational E-step: Estimate the expected values in (55)~(65), use the current distributions over the model parameters.
 - 6: The variational M-step: Update the variational solutions for each factor by (32)~(37) using the current values of the moments.
 - 7: **until** Convergence criteria is reached.
 - 8: Compute the expected value of λ_j as $\langle \lambda_j \rangle = \theta_j / (\theta_j + \vartheta_j)$ and substitute it into (14) to obtain the estimated values of the mixing coefficients π_j .
 - 9: Compute the expected value of γ_k as $\langle \gamma_k \rangle = \rho_k / (\rho_k + \varpi_k)$ and substitute it into (15) to obtain the estimated values of the mixing coefficients η_k .
 - 10: Calculate the expected values of the features saliencies by $\langle \epsilon_l \rangle = \xi_1^* / (\xi_1^* + \xi_2^*) = (\xi_1 + \sum_{i=1}^N \langle \phi_{il} \rangle) / (\xi_1 + \xi_2 + N)$.
 - 11: Detect the optimal number of components M and K by eliminating the components with small mixing coefficients close to 0.
-

4. Experimental Results

In this section, we evaluate the effectiveness of the proposed variational infinite GD mixture model with feature selection (*InFsGD*) through synthetic data and two challenging applications namely unsupervised image categorization and image annotation and retrieval. In all our experiments, we initialize the truncation levels M and K to 15 and 10, respectively. The initial values of hyperparameters u, p, g and s of the Gamma priors are set to 1, and v, q, h, t are set to 0.01. The hyperparameters a, b, c and d are set to 1, while ξ_1 and ξ_2 are set to 0.1. Our simulations have supported these specific choices.

Table 1: Parameters of the generated data sets. N denotes the total number of elements, N_j denotes the number of elements in cluster j . $\alpha_{j1}, \alpha_{j2}, \beta_{j1}, \beta_{j2}$ and π_j are the real parameters. $\hat{\alpha}_{j1}, \hat{\alpha}_{j2}, \hat{\beta}_{j1}, \hat{\beta}_{j2}$ and $\hat{\pi}_j$ are the estimated parameters by the proposed algorithm.

	N_j	j	α_{j1}	β_{j1}	α_{j2}	β_{j2}	π_j	$\hat{\alpha}_{j1}$	$\hat{\beta}_{j1}$	$\hat{\alpha}_{j2}$	$\hat{\beta}_{j2}$	$\hat{\pi}_j$
Data set 1	200	1	10	15	21	12	0.50	10.12	14.59	20.38	11.73	0.501
($N = 400$)	200	2	25	18	35	40	0.50	23.67	18.65	36.18	41.26	0.499
Data set 2	200	1	10	15	21	12	0.25	9.81	15.89	20.51	12.10	0.253
($N = 800$)	200	2	25	18	35	40	0.25	25.77	18.32	36.03	41.68	0.249
	400	3	18	35	10	25	0.50	17.35	34.29	10.72	26.65	0.498
Data set 3	200	1	10	15	21	12	0.25	10.09	15.57	21.33	11.54	0.247
($N = 800$)	200	2	25	18	35	40	0.25	24.13	17.28	35.15	38.66	0.251
	200	3	18	35	10	25	0.25	18.61	34.19	9.71	25.08	0.248
	200	4	33	27	45	13	0.25	31.95	26.83	43.89	12.27	0.254
Data set 4	200	1	10	15	21	12	0.20	9.34	14.50	20.18	12.35	0.197
($N = 1000$)	200	2	25	18	35	40	0.20	26.07	18.16	34.49	39.12	0.199
	200	3	18	35	10	25	0.20	17.31	36.53	10.76	24.22	0.203
	200	4	33	27	45	13	0.20	31.52	26.35	47.03	13.98	0.204
	200	5	20	10	42	38	0.20	19.88	10.94	41.14	36.67	0.197

4.1. Synthetic data

The purpose of the synthetic data is to investigate the accuracy of the proposed algorithm in terms of parameters estimation and model selection. The performance of the *InFsGD* was evaluated through quantitative analysis on four ten-dimensional (two relevant features and eight irrelevant features) synthetic data. The relevant features were generated in the transformed space from mixtures of Beta distributions with well-separated components, while irrelevant ones were from mixtures of overlapped components. Table 1 illustrates the real and estimated parameters of the distributions representing the relevant features for each data set using the proposed algorithm. According to this table, the parameters of the model, representing relevant features, and its mixing coefficients are accurately estimated by the *InFsGD*. Although we do not show the estimated values of the parameters of the mixture models representing irrelevant features (the eight remaining features), accurate results (in terms of both parameters estimation and model selection) were obtained by adopting the proposed algorithm.

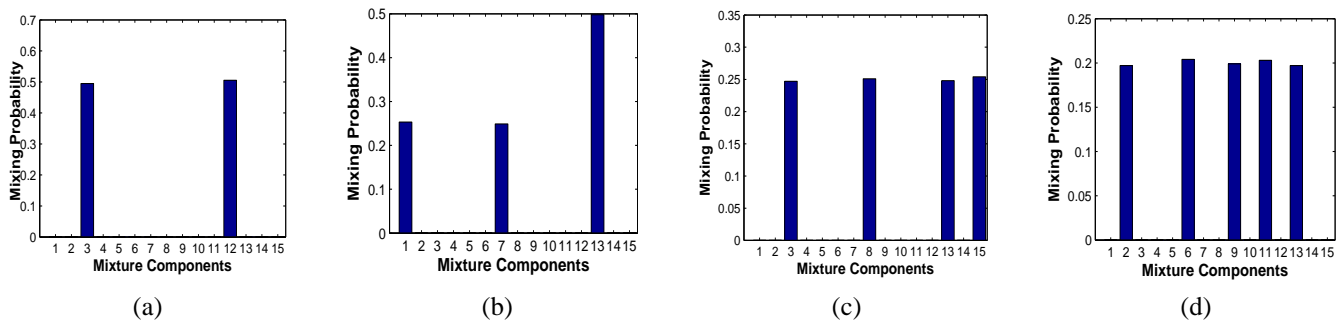


Figure 3: Mixing probabilities of components, π_j , found for each synthetic data set after convergence. (a) Data set 1, (b) Data set 2, (c) Data set 3, (d) Data set 4.

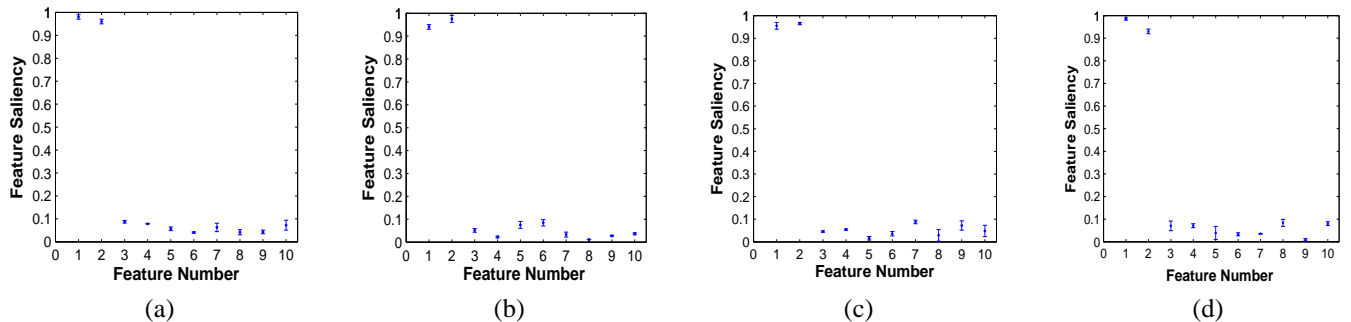


Figure 4: Features saliencies for synthetic data sets with one standard deviation over ten runs. (a) Data set 1, (b) Data set 2, (c) Data set 3, (d) Data set 4.

Figure 3 shows the estimated mixing coefficients of the mixture components, in each data set, after convergence. By removing the components with very small mixing coefficients (close to 0) in each data set, we obtain the correct number of components for the mixtures representing relevant features. Furthermore, we present the results of the features saliencies of all the 10 features for each data set over ten runs in Figure 4. It obviously shows that features 1 and 2 have been assigned a high degree of relevance, which matches the ground-truth. Furthermore, we have tested the numerical complexity of the proposed variational algorithm, in terms of overall computation time and number of iterations before convergence. The corresponding results are shown in Table 2.

4.2. Visual Scenes Categorization

In this experiment, a challenging problem namely images categorization is highlighted. It is a fundamental task in vision and has recently drawn considerable interest and has been successfully applied in various applications such as the automatic understanding of images, object recognition, image databases browsing and content-based images

Table 2: Run time (in seconds) and number of iterations required before convergence using the proposed algorithm.

Data set	Run time	No. iterations
1	6.27	362
2	11.51	419
3	12.35	433
4	16.82	467

suggestion and retrieval [59, 60]. As the majority of computer vision tasks, a central step for accurate images categorization is the extraction of good descriptors (i.e. discriminative and invariant at the same time) to represent these images. Recently local descriptors have been widely and successfully used [61, 62] mainly via the bag of visual words approach [63, 64, 65] which has allowed the development of many models inspired from text analysis such as the probabilistic latent semantic analysis (pLSA) model [66]. Recently, it has been shown that the performance of visual words-based approaches to images categorization can be significantly improved by adopting multiple image segmentations instead of considering the entire image as a way to utilize visual grouping cues to generate groups of related visual words [65, 67].

The methodology that we have adopted for categorizing images can be summarized as follows: First, we compute multiple candidate segmentations for each image in the collection using Normalized Cuts [68]¹. Following that, Gradient location-orientation histogram (GLOH) descriptors [69] are extracted from each image using the Hessian-Laplace region detector [70]². Note that, the GLOH descriptor is an extension of the SIFT descriptor, and is shown to outperform SIFT [69]. PCA is then used to reduce the dimensionality to 128. Next, a visual vocabulary \mathcal{V} is constructed by quantizing these feature vectors into visual words using K -means algorithm and each image is then represented as a frequency histogram over the visual words. Based on our experiments, the optimal performance can be obtained when $\mathcal{V} = 800$. Then, we apply the pLSA model to the bag of visual words representation which allows the description of each image as a D -dimensional vector of proportions where D is the number of aspects (or learnt topics). Finally, we employ the proposed *InFsGD* as a classifier to categorize images by assigning each test image to the class which has the highest posterior probability according to Bayes' decision rule.

In our experiment, two challenging data sets have been employed: First, we have considered four object classes from the Caltech data set [71] which include: "airplane", "face", "car", and "motorbike"; The second data set is the UIUC sports event data set [72] containing 8 categories of sports scenes: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snow boarding (190 images), croquet (236 images), sailing (190 images), and rock climbing (194 images). Thus, the data set contains 1,579 images in total. Sample images from these two data sets are displayed in Figures 5 and 6. Each of the two data sets is randomly divided into two halves: one for training (constructing the visual words) and the other for testing. We evaluated the performance of the proposed algorithm by running it 20 times. For comparison, we have applied the infinite GD mixture model using Gibbs sampling algorithm (denoted as *GiInGD*) proposed in [24] and four other mixture models that are learned using variational inference: the finite GD mixture model with feature selection (*FsGD*), the infinite GD mixture model without feature selection (*InGD*), the infinite Gaussian mixture model (*InGau*) proposed in [47] and the Gaussian mixture model with feature selection (*FsGau*) as learned in [42].

The performance of the proposed approach for image categorization was first evaluated on the Caltech data set. First, multiple segmentations for each image are performed. Some sample segments for images from each category in this data set are shown in Figure 7. The categorization accuracies using the different tested approaches are presented in Table 3. According to the results in this table, the proposed *InFsGD* provides the best performance among the tested algorithms in terms of the highest classification rate and the most accurately estimation of the number of categories. It is noteworthy that the variational approach (*InGD*) provides comparable performance as the one using Gibbs sampling (*GiInGD*) according to Table 3. However, the major advantage of using variational inference algorithm is its computational efficiency. According to our experiment, *InGD* is almost four times faster than *GiInGD* for this application.

¹Source code of the Normalized Cuts segmentation is available at: <http://www.seas.upenn.edu/~timothee/software/ncut/ncut.html>

²Source code: <http://www.robots.ox.ac.uk/~vgg/research/affine/>



Figure 5: Sample images from the four categories of the Caltech data set.



Figure 6: Sample images from the UIUC sports event data set.

Additionally, the number of components for the mixture model representing irrelevant features was estimated as 2 using the proposed *InFsGD*. Furthermore, we have tested the evolution of the classification accuracy with different number of aspects as shown in Figure 8 (a). Based on this figure, the highest classification accuracy can be obtained when we set the number of aspects to 40. The corresponding feature saliencies of the 40 aspects obtained by *InFsGD* are illustrated in Figure 8 (b). As shown in this figure, it is clear that the features have different relevance degrees and then contribute differently to images categorization. Next, we evaluate the effectiveness and the efficiency of the

Table 3: The average classification accuracy and the number of categories (\hat{M}) computed by different algorithms for the Caltech data set.

	<i>InFsGD</i>	<i>FsGD</i>	<i>InGD</i>	<i>GiInGD</i>	<i>InGau</i>	<i>FsGau</i>
\hat{M}	3.90	3.75	3.85	3.85	3.80	3.70
Accuracy (%)	90.21	88.64	88.03	88.59	84.19	81.75

proposed approach on categorizing the UIUC sports event data set. The confusion matrix for this data set calculated

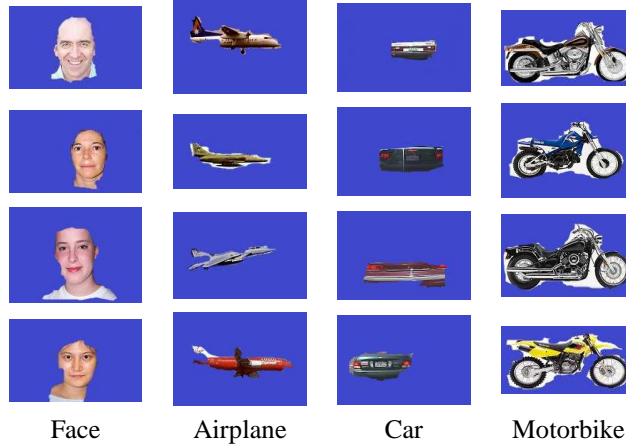


Figure 7: Sample segmentation results from the four categories of the Caltech data set.

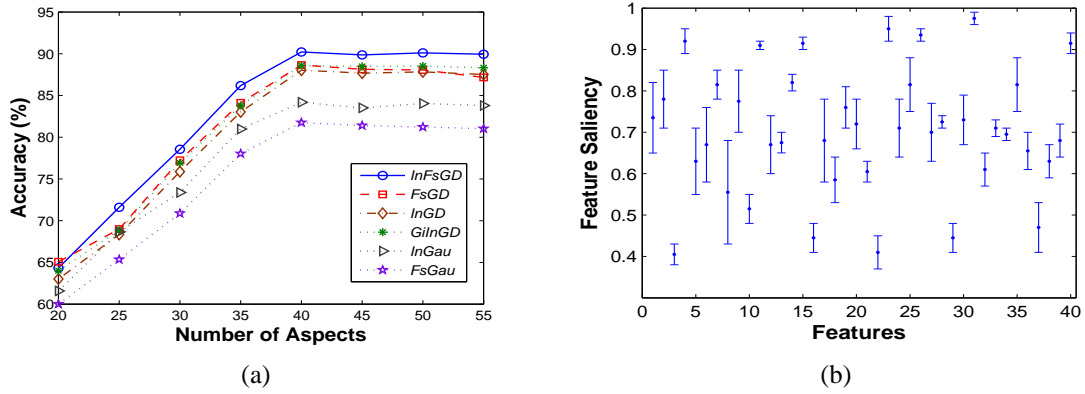


Figure 8: Caltech data set: (a) Classification accuracy vs. the number of aspects; (b) Feature saliency for each aspect.

by the *InFsGD* is shown in Figure 9. The average categorization accuracies and the average numbers of categories obtained by the different algorithms are shown in Table 4. As we can see from this table, our algorithm provides the highest accuracy while detecting the number of categories more accurately. In this experiment, the optimal number

Table 4: The average classification accuracy and the number of categories (\hat{M}) computed by different algorithms for the UIUC sports data set.

	<i>InFsGD</i>	<i>FsGD</i>	<i>InGD</i>	<i>GilnGD</i>	<i>InGau</i>	<i>FsGau</i>
\hat{M}	7.85	7.60	7.75	7.80	7.70	7.60
Accuracy (%)	74.13	72.08	71.76	71.83	68.37	65.51

of aspects is 50 as shown in Figure 10 (a). The saliencies of the 50 aspects calculated by *InFsGD* are illustrated in Figure 10 (b). Obviously, different features are assigned with different degrees of importance. For instance, there are six features (features number 10, 16, 23, 31, 33 and 46) that have saliencies lower than 0.5, and then provide less contribution to clustering. By contrast, nine features (features number 2, 9, 12, 17, 28, 38, 41, 48 and 49) have high relevance degrees with saliencies greater than 0.9. Finally, we have compared the proposed method with two traditional and widely used classifiers: k-nearest neighbors (*KNN*) and support vector machines (*SVM*). These two classifiers have shown their effectiveness in scene categorization task with bag of visual words framework previously in [64, 73]. In our work, an Euclidean distance function was used for *KNN* (with $K = 10$) as in [64] while an Euclidean exponential kernel was adopted for *SVM* as in [73]. The corresponding results are shown in Table 5. According to

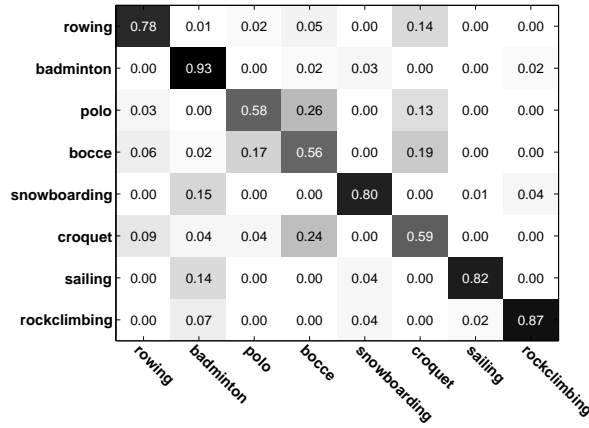


Figure 9: Confusion matrix obtained by *InFsGD* for the UIUC sports data set.

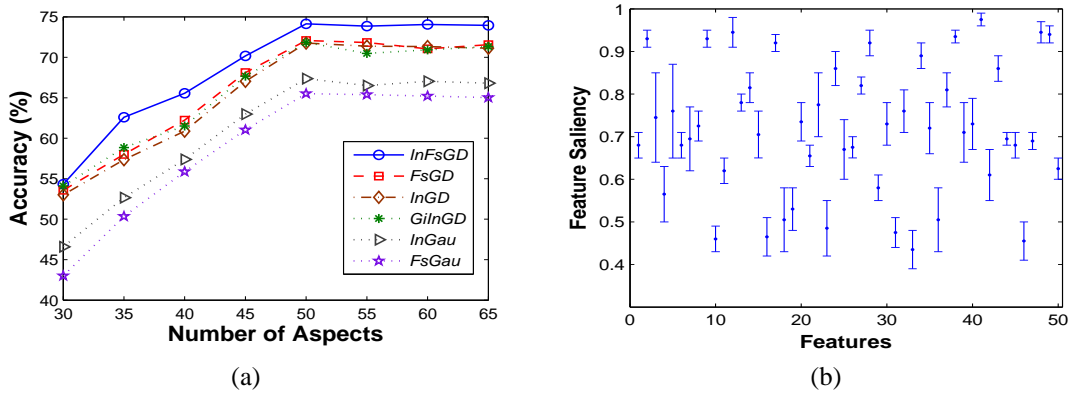


Figure 10: UIUC sports data set: (a) Classification accuracy vs. the number of aspects; (b) Feature saliency for each aspect.

the results, it is clear that our method outperforms both *KNN* and *SVM* in terms of classification accuracy for both data sets. This is due to the fact that either *KNN* or *SVM* considers all aspects with “equal” importance. However, as shown in Figures 8 (b) and 10 (b), these aspects may contribute with different degrees in discriminating the image categories. Thus, higher classification accuracies can be obtained by using *InFsGD* which allows to identify and discard the aspects without discrimination power.

Table 5: The average classification accuracy (%) computed by different algorithms for the Caltech and the UIUC sports data sets.

Data set	<i>InFsGD</i>	<i>KNN</i>	<i>SVM</i>
Caltech	90.21	84.32	88.46
UIUC sports	74.13	68.55	72.38

4.3. Image Auto-Annotation

4.3.1. Methodology

Many images carrying extremely rich information are now archived in large databases. A challenging problem is then to automatically analyze, organize, index, browse and retrieve these images. A lot of approaches have been proposed to address this problem. In particular, semantic image understanding and auto-annotation have been the topic of extensive research in the past [74, 75, 76, 77, 78, 79, 80]. The main goal is to extract high-level semantic features in addition to low level features to bridge the gap between them and to enhance visual scenes interpretation abilities [81, 82, 83]. Automatic annotation approaches can be divided into two main groups of approaches [84, 85]. The first group deals directly with the annotation problem by providing labels to the complete image or its different regions (see, for instance, [81, 82]). The second group tackles this problem via two independent steps where the first step categorizes the images and the second one affects labels to them using the top ranked categories (see, for instance, [76, 85]). Approaches in this second group have shown promising results recently. Thus, the goal of this subsection is to develop an annotation-driven image retrieval approach, based on the work in [85], via categorization results obtained with the proposed *InFsGD* in a bag of visual key words representation. Our aim is to build an efficient annotation-retrieval approach to handle the problem of image search under three challenging scenarios as stated in [85]: 1) use a tagged image or a set of keywords as query to search images on the untagged portion of a partially tagged image database; 2) use an untagged image as query to search images on the tagged portion of a partially tagged image database; 3) use an untagged image as query to search images on an untagged image database. The methodology that we have adopted for this experiment can be divided into three sequential steps namely: images categorization, annotation, and retrieval.

In the categorization stage, the proposed *InFsGD* is integrated with the pLSA model to categorize images through a bag of key visual words representation. First, interest points are detected using the Difference-of-Gaussian (DoG) detector [70]. Then, we use PCA-SIFT descriptor³ [86], computed on detected keypoints of all images and resulting on 36-dimensional vector for each keypoint. Subsequently, the *K*-Means algorithm is used to construct a visual vocabulary by quantizing these PCA-SIFT vectors into visual words. In our experiments, we set the vocabulary size to 1000. Each image is then represented as a frequency histogram over the visual words. Then, the pLSA model is applied to the obtained histograms to represent each image by a 50-dimensional proportional vector where 50 is the number of latent aspects. Finally, our *InFsGD* is deployed to cluster the images. The categorization results in the previous stage are exploited to perform image annotation. Here, we follow an approach proposed in [85] which considers the problem of image annotation from three phases: 1) the frequency of occurrence of potential tags based on the categorization results; 2) saliency of the given tags; 3) the congruity of a word among all the candidate tags. Assume that we have a training image data set that contains several categories. Each category is annotated by 4 to 5 tags where common tags may appear in different categories. At the beginning, we collect all the tags from each category. The total number of categories in the data set is denoted as C and the number of categories that have each unique tag t is represented as $F(t)$. Then, tag saliency can be evaluated similarly as for inverse document frequency in the field of document retrieval. For a test image, a ranked list of predicted categories is generated according to the Bayes' decision rule in the classification. Then, the top 5 predicted categories are chosen and the union of all involved unique tags denoted as $U(I)$ forms the set of candidate tags. Thus, we define $f(t|I)$ as the frequency of the occurrence of each unique tag t among the top 5 predicted categories. We follow the idea proposed in [85] to determine the word congruity using WordNet [87] with the Leacock and Chowdrow measure [88]. WordNet is a large lexical database of English which groups English words into sets of cognitive synonyms called synsets. Hence, the congruity for a candidate tag t can be calculated by [85]:

$$G(t|I) = \frac{d_{tot}(I)}{d_{tot}(I) + |U(I)| \sum_{x \in U(I)} d_{LCH}(x, t)} \quad (66)$$

We adopt the same settings for d_{LCH} and r_{LCH} as in [85], such that the distance between two tags t_1 and t_2 is: $d_{LCH}(t_1, t_2) = \exp(-r_{LCH}(t_1, t_2) + 3.584) - 1$. In addition, $d_{tot}(I)$ evaluates the pairwise semantic distance among all candidate tags and is defined as: $d_{tot}(I) = \sum_{x \in U(I)} \sum_{y \in U(I)} d_{LCH}(x, y)$. By having all the three annotation factors on

³Source code of PCA-SIFT: <http://www.cs.cmu.edu/~yke/pcasift>

hand, we can compute the overall score for a candidate tag as

$$A(t|I) = a_1 f(t|I) + \frac{a_2}{\ln C} \ln\left(\frac{C}{1 + F(t)}\right) + a_3 G(t|I) \quad (67)$$

where $a_1 + a_2 + a_3 = 1$ represents the degree of importance of the three factors. Then, a tag t is chosen for annotation only if its score is within the top ε percentile among the candidate tags. According to our experimental results, we set $a_1 = 0.5$, $a_2 = 0.2$, $a_3 = 0.3$, and $\varepsilon = 0.7$. For retrieving images, we use automatic annotation and the WordNet-based bag of words distances as introduced in [85]. The core idea is that if tags were missing in the query image or in our database, automatic annotation is then performed and the bag of words distances between query image tags and the database tags are calculated. This distance is used to rank the degree of relevance of the images in the database and then to perform images search accordingly (more details and discussions can be found in [85]).

4.3.2. Results

We test out approach using a subset of LabelMe data set [89] which contains both class labels and annotations. First, we use the LabelMe Matlab toolbox⁴ to obtain images online from 8 outdoor scene classes: “highway”, “inside city”, “tall building”, “street”, “forest”, “coast”, “mountain” and “open country”. We randomly choose 200 images from each categories. Thus, we have 1600 images in total. Each category is associated with 4-5 tags. We randomly divide the data set into two partitions: one for training, the other for testing. First, we have performed categorization using the proposed *InFsGD* with bag of visual key words representation as described previously. We compare our approach with other five well-defined approaches: the variational infinite GD mixture model without feature selection (*InGD*), the infinite GD mixture model using Gibbs sampling algorithm (*GiInGD*), the variational infinite Gaussian mixture model (*InGau*), the combination of a structure-composition model and a Gaussian mixture model (we denote it as *SC-GM*) as proposed in [85] and the variational Gaussian mixture model with feature selection (*FsGau*). The categorization result of the 8 outdoor scene images is illustrated in Table 6. According to this table, we can observe that the proposed *InFsGD* outperforms other five approaches in terms of the highest classification accuracy rate (75.1%). The obtained result from the categorization is then exploited by the annotation stage. The performance of annotation

Table 6: The average classification accuracy computed by different algorithms.

Method	Accuracy (%)
<i>InFsGD</i>	75.1
<i>InGD</i>	74.7
<i>GiInGD</i>	74.8
<i>InGau</i>	73.6
<i>SC-GM</i>	71.8
<i>FsGau</i>	70.2

is evaluated by precision and recall which are defined in the standard way: the annotation precision for a keyword is defined as the number of tags correctly predicted divided by the total number of predicted tags. The annotation recall is defined as the number of tags correctly predicted, divided by the number of tags in the ground-truth annotation. In our experiments, the average number of tags generated for each test image is 4.05. Table 7 shows the performance evaluation of the automatic annotation approach according to the categorization result obtained by using different methods. It is clear that, annotation with the categorization result obtained by *InFsGD* provides the best performance. Table 8 presents some examples of the annotations produced by using *InFsGD* categorization method. In the last step, we perform image retrieval under the three scenarios as described in the previous subsection. For the first scenario in which the database is not tagged and query may either be keywords or tagged image, the retrieval is performed by first automatically annotating the database through categorization and annotation steps. Then, image retrieval is

⁴<http://labelme.csail.mit.edu/>

Table 7: Performance evaluation of the automatic annotation system based on different categorization methods.

Method	Mean Precision (%)	Mean Recall (%)
<i>InFsGD</i>	31.5	43.6
<i>InGD</i>	30.4	42.3
<i>GiInGD</i>	30.3	42.5
<i>InGau</i>	29.8	40.2
<i>SC-GM</i>	27.1	38.7
<i>FsGau</i>	26.3	36.8

Table 8: Sample annotation results by using *InFsGD* classification method.









				
Our labels	car, road, mountain	car, sidewalk, window	sky, building, tree	human, car, tree
LabelMe labels	truck, car, sky, road, mountain	building, car, window, sidewalk, human	building, tree, car, sky	person, car, sidewalk, building, tree
				
Our labels	sea water, tree, sky	sand, tree, sea water	forest, sky, cloud	cloud, field, mountain, tree
LabelMe labels	tree, forest, mountain, cloud, sky	sea water, sand, sky, cloud	mountain, sky, field, tree	sky, sand, field, mountain, car

Table 9: The comparison of image retrieval performances.

Method	Scenario 1		Scenario 2		Scenario 3	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
<i>InFsGD</i>	51.5	58.9	45.3	50.2	47.5	56.6
<i>InGD</i>	49.7	56.6	42.5	49.3	46.6	54.1
<i>GiInGD</i>	49.9	56.5	42.8	49.7	46.5	54.3
<i>InGau</i>	48.6	56.3	41.4	48.7	45.9	52.8
<i>SC-GM</i>	46.2	55.7	38.6	45.6	41.7	53.5
<i>FsGau</i>	43.8	52.1	37.1	43.4	38.3	51.0

performed according to the bag of words distances between query tags and our annotation. In this experiment, we use 40 pairs of query words that are randomly chosen from all the candidate tags. In the second scenario, the database is tagged and the query is an untagged image. Thus, the first step is to automatically annotate the query image. Then, the database is ranked according to the bag of words distances. In the third scenario, neither the image database nor the query is tagged. Therefore, both the image database and the query images have to be annotated automatically first. Subsequently, image retrieval is applied once again using the bag of words distance evaluation. We choose 100 images randomly as the set of query images in this experiment. The performance of semantic retrieval was evaluated by measuring precision and recall. In this case, precision is defined as the proportion of retrieved images that are relevant, and recall denotes the proportion of relevant images that are retrieved. An image is considered relevant if there is an overlap between the original tags of the query image or query word and the original tags of the retrieved image. Since categorization is the baseline of our annotation-driven image retrieval approach. We have also tested the impact of using different categorization algorithms on annotation-driven image retrieval performance and illustrates the corresponding result in Table 9 on retrieving the top 10 relevant images. As we can observed form this table, using *InFsGD* as the categorization method provides the best performance for all three scenarios which indicates that the categorization algorithm is a significant influence factor for the annotation-driven image retrieval scheme that we have applied.

5. Conclusion

Until recently, feature selection approaches based on mixture models were almost exclusively considered in the finite case. The work proposed in this paper is motivated by an attempt to overcome this limitation via the extension of the simultaneous clustering and feature selection approach based on finite generalized Dirichlet mixture models, previously proposed in [1], to the infinite case via Dirichlet processes with a stick-breaking representation. The proposed technique drives much of its power from the flexibility of the generalized Dirichlet mixture, the high generalization accuracy of Dirichlet processes, and the advantages of the variational Bayesian framework that we have developed to learn our model. Our method has been successfully tested in several scenarios and our experimental results using synthetic data and real-world applications namely visual scenes categorization, auto-annotation and retrieval have shown advantages derived from its adoption. The model developed in this paper is also applicable to many other problems which involve high-dimensional data clustering such as gene microarray data sets analysis, text clustering and retrieval, and object recognition.

Appendix A. Proof of Equations (32)~(37)

Based on (28), the general expression for the variational solution $Q_s(\Theta_s)$ can be written as

$$\ln Q_s(\Theta_s) = \langle \ln p(\mathcal{X}, \Theta) \rangle_{i \neq s} + \text{const.} \quad (\text{A.1})$$

where any terms that are independent of $Q_s(\Theta_s)$ are absorbed into the additive constant. Next, we need to calculate the logarithm of the joint distribution (26) as

$$\begin{aligned} \ln p(\mathcal{X}, \Theta) &= \sum_{i=1}^N \sum_{j=1}^{\infty} Z_{ij} \left\{ \sum_{l=1}^D \phi_{il} \left[\ln \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} + (\alpha_{jl} - 1) \ln X_{il} + (\beta_{jl} - 1) \ln(1 - X_{il}) \right] + \sum_{l=1}^D \sum_{k=1}^{\infty} (1 - \phi_{il}) W_{ikl} \left[\ln \frac{\Gamma(\sigma_{kl} + \tau_{kl})}{\Gamma(\sigma_{kl})\Gamma(\tau_{kl})} \right. \right. \\ &+ (\sigma_{kl} - 1) \ln X_{il} + (\tau_{kl} - 1) \ln(1 - X_{il}) \left. \left. \right] \right\} + \sum_{i=1}^N \sum_{j=1}^{\infty} Z_{ij} \left[\ln \lambda_j + \sum_{s=1}^{j-1} \ln(1 - \lambda_s) \right] + \sum_{j=1}^{\infty} [\ln \psi_j + (\psi_j - 1) \ln(1 - \lambda_j)] + \sum_{j=1}^{\infty} [(a_j - 1) \ln \psi_j - b_j \psi_j] \\ &+ \sum_{i=1}^N \sum_{k=1}^{\infty} \sum_{l=1}^D W_{ikl} \left[\ln \gamma_k + \sum_{s=1}^{k-1} \ln(1 - \gamma_s) \right] + \sum_{k=1}^{\infty} [\ln \varphi_k + (\varphi_k - 1) \ln(1 - \varphi_k)] + \sum_{k=1}^{\infty} [(c_k - 1) \ln \varphi_k - d_k \varphi_k] + \sum_{i=1}^N \sum_{l=1}^D [\phi_{il} \ln \epsilon_{l_1} + (1 - \phi_{il}) \ln \epsilon_{l_2}] \\ &+ \sum_{l=1}^D [(\xi_1 - 1) \ln \epsilon_{l_1} + (\xi_2 - 1) \ln \epsilon_{l_2}] + \sum_{j=1}^{\infty} \sum_{l=1}^D [(u_{jl} - 1) \ln \alpha_{jl} - v_{jl} \alpha_{jl} + (p_{jl} - 1) \ln \beta_{jl} - q_{jl} \beta_{jl}] \\ &+ \sum_{k=1}^{\infty} \sum_{l=1}^D [(g_{kl} - 1) \ln \sigma_{kl} - h_{kl} \sigma_{kl} + (s_{kl} - 1) \ln \tau_{kl} - t_{kl} \tau_{kl}] + \text{const.} \end{aligned} \quad (\text{A.2})$$

In the following variational inference process, we truncate the stick-breaking representation for the infinite GD mixture model at a value of M , and the infinite Beta mixture model for the irrelevant features is truncated at a level of K .

Appendix A.1. Variational Solution to $Q(\vec{\phi})$

We can compute the logarithm of the variational factor $Q(\phi_{il})$ as

$$\begin{aligned} \ln Q(\phi_{il}) = \langle \ln(p(\mathcal{X}, \Theta)) \rangle_{\Theta \neq \phi_{il}} = \phi_{il} \left\{ \sum_{j=1}^M \langle Z_{ij} \rangle [\mathcal{R}_{jl} + (\bar{\alpha}_{jl} - 1) \ln X_{il} + (\bar{\beta}_{jl} - 1) \ln(1 - X_{il})] + \langle \ln \epsilon_{l_1} \rangle \right\} \\ + (1 - \phi_{il}) \left\{ \sum_{k=1}^K \langle W_{ikl} \rangle [\mathcal{F}_{kl} + (\bar{\sigma}_{kl} - 1) \ln X_{il} + (\bar{\tau}_{kl} - 1) \ln(1 - X_{il})] \langle \ln \epsilon_{l_2} \rangle \right\} + \text{const.} \end{aligned} \quad (\text{A.3})$$

where

$$\bar{\alpha} = \langle \alpha \rangle, \quad \bar{\beta} = \langle \beta \rangle, \quad \bar{\sigma} = \langle \sigma \rangle, \quad \bar{\tau} = \langle \tau \rangle \quad (\text{A.4})$$

and we define

$$\mathcal{R}_{jl} = \left\langle \ln \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} \right\rangle_{\alpha_{jl}, \beta_{jl}}, \quad \mathcal{F}_{kl} = \left\langle \ln \frac{\Gamma(\sigma_{kl} + \tau_{kl})}{\Gamma(\sigma_{kl})\Gamma(\tau_{kl})} \right\rangle_{\sigma_{kl}, \tau_{kl}} \quad (\text{A.5})$$

Note that the expectations in (A.5) are analytically intractable, thus, the standard variational inference can not be applied directly. To tackle this problem, we can apply a lower bound approximation, such as second-order Taylor series expansion, to the intractable function to obtain a closed-form expression [51, 56]. In our work, we adopt the second-order Taylor series expansion to approximate \mathcal{R}_{jl} and \mathcal{F}_{kl} using $\tilde{\mathcal{R}}_{jl}$ (43) and $\tilde{\mathcal{F}}_{kl}$ (44) as proposed in [56]. By substituting the lower bounds (43) and (44) into (A.3), we then obtain

$$\begin{aligned} \ln Q(\phi_{il}) = \langle \ln(p(\mathcal{X}, \Theta)) \rangle_{\Theta \neq \phi_{il}} = \phi_{il} \left\{ \sum_{j=1}^M \langle Z_{ij} \rangle [\tilde{\mathcal{R}}_{jl} + (\bar{\alpha}_{jl} - 1) \ln X_{il} + (\bar{\beta}_{jl} - 1) \ln(1 - X_{il})] + \langle \ln \epsilon_{l_1} \rangle \right\} \\ + (1 - \phi_{il}) \left\{ \sum_{k=1}^K \langle W_{ikl} \rangle [\tilde{\mathcal{F}}_{kl} + (\bar{\sigma}_{kl} - 1) \ln X_{il} + (\bar{\tau}_{kl} - 1) \ln(1 - X_{il})] \langle \ln \epsilon_{l_2} \rangle \right\} + \text{const.} \end{aligned} \quad (\text{A.6})$$

We can find that (A.6) has the same logarithmic form of (20) except for the normalization constant. Therefore, we can acquire the variational solution as

$$Q(\vec{\phi}) = \prod_{i=1}^N \prod_{l=1}^D f_{il}^{\phi_{il}} (1 - f_{il})^{(1 - \phi_{il})} \quad (\text{A.7})$$

where f_{il} is defined in (38). Then, from the Bernoulli distribution $Q(\vec{\phi})$ (A.7), it is straightforward to have

$$\langle \phi_{ij} \rangle = f_{ij} \quad \text{and} \quad \langle 1 - \phi_{ij} \rangle = 1 - f_{ij} \quad (\text{A.8})$$

Appendix A.2. Variational Solution to $Q(\mathcal{Z})$

The logarithm of the variational factor $Q(Z_{ij})$ is calculated as

$$\ln Q(Z_{il}) = Z_{ij} \left\{ \sum_{l=1}^D \langle \phi_{il} \rangle [\tilde{\mathcal{R}}_{jl} + (\bar{\alpha}_{jl} - 1) \ln X_{il} + (\bar{\beta}_{jl} - 1) \ln(1 - X_{il})] + \langle \ln \lambda_j \rangle + \sum_{s=1}^{j-1} \langle \ln(1 - \lambda_s) \rangle \right\} + \text{Const.} \quad (\text{A.9})$$

where we have substituted $\tilde{\mathcal{R}}_{jl}$ for \mathcal{R}_{jl} . By analyzing (A.9), it is obvious that the variational solution to $Q(\mathcal{Z})$ has the logarithmic form of (12) except for the normalization constant. Therefore, we can rewrite (A.9) as

$$\ln Q(\mathcal{Z}) = \sum_{i=1}^N \sum_{j=1}^M Z_{ij} \ln \tilde{r}_{ij} + \text{const.} \quad (\text{A.10})$$

where \tilde{r}_{ij} is defined in (39). By taking the exponential of both sides of (A.10), we then have

$$Q(\mathcal{Z}) \propto \prod_{i=1}^N \prod_{j=1}^M \tilde{r}_{ij}^{Z_{ij}} \quad (\text{A.11})$$

Since Z_{ij} are binary and $\sum_{j=1}^M Z_{ij} = 1$, (A.11) can be normalized as

$$Q(\mathcal{Z}) = \prod_{i=1}^N \prod_{j=1}^M r_{ij}^{Z_{ij}} \quad (\text{A.12})$$

where r_{ij} is given in (38). Since r_{ij} is nonnegative and sum to one, for the multinomial distribution $Q(\mathcal{Z})$ we can obtain

$$\langle z_{ij} \rangle = r_{ij} \quad (\text{A.13})$$

where r_{ij} are playing the role of responsibilities as in the conventional EM algorithm.

Appendix A.3. Variational Solution to $Q(\vec{\lambda})$

For the variational factor $Q(\vec{\lambda})$, its logarithm form is obtained by

$$\ln Q(\lambda_j) = \ln \lambda_j \sum_{i=1}^N \langle Z_{ij} \rangle + \ln(1 - \lambda_j) \left(\sum_{i=1}^N \sum_{s=j+1}^M \langle Z_{is} \rangle + \langle \psi_j \rangle - 1 \right) + \text{Const.} \quad (\text{A.14})$$

We can observe that (A.14) has the logarithmic form of a Beta distribution as its conjugate prior distribution (17). By taking the exponential of its both sides, we obtain

$$Q(\vec{\lambda}) = \prod_{j=1}^M \text{Beta}(\lambda_j | \theta_j, \vartheta_j) \quad (\text{A.15})$$

where θ_j and ϑ_j are defined in (51).

Appendix A.4. Variational Solution to $Q(\vec{\psi})$

The logarithm form of the variational factor $Q(\vec{\beta})$ is given by

$$\ln Q(\psi_j) = \ln \psi_j a_j + \psi_j (\langle \ln(1 - \lambda_j) \rangle - b_j) + \text{Const.} \quad (\text{A.16})$$

By taking the exponential of the both sides of (A.16), we can obtain

$$Q(\vec{\psi}) = \prod_{j=1}^M \mathcal{G}(\psi_j | a_j^*, b_j^*) \quad (\text{A.17})$$

Thus, the optimal solutions to the hyper-parameters a_j^* and b_j^* can be calculated as

$$a_j^* = a_j + 1, \quad b_j^* = b_j - \langle \ln(1 - \lambda_j) \rangle \quad (\text{A.18})$$

Appendix A.5. Variational Solution to $Q(\mathcal{W})$, $Q(\vec{\gamma})$ and $Q(\vec{\varphi})$

We can calculate the logarithm of the variational factor $Q(W_{ikl})$ as

$$\ln Q(W_{ikl}) = W_{ikl} \left\{ \langle 1 - \phi_{il} \rangle [\tilde{\mathcal{F}}_{kl} + (\tilde{\sigma}_{kl} - 1) \ln X_{il} + (\tilde{\tau}_{kl} - 1) \ln(1 - X_{il})] + \langle \ln \gamma_k \rangle + \sum_{s=1}^{k-1} \langle \ln(1 - \gamma_s) \rangle \right\} + \text{Const.} \quad (\text{A.19})$$

where we have substituted $\tilde{\mathcal{F}}_{kl}$ for \mathcal{F}_{kl} . By analyzing (A.19), we can obtain the variational solution to $Q(\mathcal{W})$ by taking the exponential of the both sides it as

$$Q(\mathcal{W}) = \prod_{i=1}^N \prod_{k=1}^K \prod_{l=1}^D m_{ikl}^{W_{ikl}} \quad (\text{A.20})$$

where m_{ikl} is given by (38).

Since $\vec{\gamma}$ has the Beta prior and $\vec{\varphi}$ has the Gamma prior distribution, the variational solutions to $Q(\vec{\gamma})$ and $Q(\vec{\varphi})$ can be derived in a similar way as for $Q(\vec{\lambda})$ and $Q(\vec{\psi})$, respectively.

Appendix A.6. Variational Solution to $Q(\vec{\epsilon})$

The logarithm of the variational factor $Q(\vec{\epsilon})$ as

$$\ln Q(\vec{\epsilon}) = \ln \epsilon_{l_1} \left(\sum_{i=1}^N \langle \phi_{il} \rangle + \xi_1 - 1 \right) + \ln \epsilon_{l_2} \left(\sum_{i=1}^N \langle 1 - \phi_{il} \rangle + \xi_2 - 1 \right) + \text{Const.} \quad (\text{A.21})$$

We can see that (A.21) has logarithmic form of a Dirichlet distribution except for the normalization constant. Then, the variational solution to $Q(\vec{\epsilon})$ can be obtained by

$$Q(\vec{\epsilon}) = \prod_{l=1}^D \text{Dir}(\vec{\epsilon}_l | \vec{\xi}^*) \quad (\text{A.22})$$

where $\vec{\xi}^*$ are defined in (54).

Appendix A.7. Variational Solution to $Q(\vec{\alpha})$, $Q(\vec{\beta})$, $Q(\vec{\sigma})$ and $Q(\vec{\tau})$

The logarithm of the variational factor $Q(\alpha_{jl})$ can be calculated as

$$\ln Q(\alpha_{jl}) = \langle \ln p(\mathcal{X}, \Theta) \rangle_{\Theta \neq \alpha_{jl}} = \sum_{i=1}^N \langle Z_{ij} \rangle \langle \phi_{il} \rangle [\mathcal{D}(\alpha_{jl}) + \alpha_{jl} \ln X_{il}] + (u_{jl} - 1) \ln \alpha_{jl} - v_{jl} \alpha_{jl} + \text{const.} \quad (\text{A.23})$$

where we have defined

$$\mathcal{D}(\alpha_{jl}) = \left\langle \ln \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} \right\rangle_{\beta_{jl}} \quad (\text{A.24})$$

Since the function $\mathcal{D}(\alpha_{jl})$ is analytically intractable, we can not perform the standard variational inference directly and (A.23) does not have the same form as the logarithm of a Gamma distribution as its conjugate prior. Thus, we approximate the function $\mathcal{D}(\alpha)$ by a non-linear approximation as proposed in [56], such that

$$\mathcal{D}(\alpha) \geq \ln \alpha \{ \Psi(\bar{\alpha} + \bar{\beta}) - \Psi(\bar{\alpha}) + \bar{\beta}' \Psi'(\bar{\alpha} + \bar{\beta}) (\langle \ln \beta \rangle - \ln \bar{\beta}) \} \bar{\alpha} \quad (\text{A.25})$$

After substituting the lower bound (A.25) back into (A.23), we then have

$$\begin{aligned} \ln Q(\alpha_{jl}) \approx & \ln \alpha_{jl} \left\{ \sum_{i=1}^N \langle z_{ij} \rangle \langle \phi_{il} \rangle [\Psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \Psi(\bar{\alpha}_{jl}) + \bar{\beta}_{jl}' \Psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) (\langle \ln \beta_{jl} \rangle - \ln \bar{\beta}_{jl})] \bar{\alpha}_{jl} + u_{jl} - 1 \right\} \\ & + \alpha_{jl} \left[\sum_{i=1}^N \langle X_{ij} \rangle \langle \phi_{il} \rangle \ln X_{il} - v_{jl} \right] + \text{Const.} \end{aligned} \quad (\text{A.26})$$

We can find that (A.26) has the logarithmic form of a Gamma distribution. By taking the exponential of both sides of (A.26), we then obtain

$$Q(\vec{\alpha}) = \prod_{j=1}^M \prod_{l=1}^D \mathcal{G}(\alpha_{jl} | u_{jl}^*, v_{jl}^*) \quad (\text{A.27})$$

The hyperparameters u_{jl}^* and v_{jl}^* can be estimated by (45) and (49).

Since $\vec{\beta}$, $\vec{\sigma}$ and $\vec{\tau}$ all have Gamma prior, it is straightforward to obtain the variational solutions to $Q(\vec{\beta})$, $Q(\vec{\sigma})$ and $Q(\vec{\tau})$ in a same way as for $Q(\vec{\alpha})$.

Acknowledgment

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors would like to thank the anonymous referees and the associate editor for their comments. The complete source code of this work is available upon request.

References

- [1] S. Boutemedjet, N. Bouguila, D. Ziou, A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (2009) 1429–1443.
- [2] H. Frigui, R. Krishnapuram, Clustering by Competitive Agglomeration, *Pattern Recognition* 30 (1997) 1109–1119.
- [3] J. G. Campbell, C. Fraley, F. Murtagh, A. E. Raftery, Linear Flaw Detection in Woven Textiles Using Model-Based Clustering, *Pattern Recognition Letters* 18 (1997) 1539–1548.
- [4] T. Lange, V. Roth, M. L. Braun, J. M. Buhmann, Stability-Based Validation of Clustering Solutions, *Neural Computation* 16 (2004) 1299–1323.
- [5] C. Ding, X. He, K-Means Clustering via Principal Component Analysis, in: *Proc. of the twenty-first international conference on Machine learning (ICML)*, ACM, 2004, pp. 29–37.
- [6] R. Ostrovsky, Y. Rabani, L. J. Schulman, C. Swamy, The Effectiveness of Lloyd-Type Methods for The K-Means Problem, in: *Proc. of the 47th Annual IEEE Symposium on Foundations of Computer Science, FOCS '06*, IEEE Computer Society, Washington, DC, USA, 2006, pp. 165–176.
- [7] M. Meila. The Uniqueness of a Good Optimum for K-Means, in: *Proc. of the 23rd International Conference on Machine Learning (ICML)*, ACM, 2006, pp. 625–632.
- [8] C. C. Aggarwal, P. S. Yu, Finding Generalized Projected Clusters in High Dimensional Spaces, in: *Proceedings of the ACM SIGMOD conference on Management of data (SIGMOD)*, ACM, 2000, pp. 70–81.
- [9] J. Bins, B. A. Draper, Feature Selection from Huge Feature Sets, in: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 2001, pp. 159–165.
- [10] X. J. Zhou, T. S. Dillon, A Statistical-Heuristic Feature Selection Criterion for Decision Tree Induction, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (1991) 834–841.
- [11] L. Xie, P. Pérez, Slightly Supervised Learning of Part-Based Appearance Models, in: *Proc. of the IEEE Workshop on Learning in Computer Vision and Pattern Recognition*, IEEE Computer Society, 2004, pp. 100–107.
- [12] M. Clyde, G. Parmigiani, B. Vidakovic, Multiple Shrinkage and Subset Selection in Wavelets, *Biometrika* 85 (1998) 391–401.
- [13] G. McLachlan, D. Peel, *Finite Mixture Models*, New York: Wiley, 2000.
- [14] E. Come, L. Oukhellou, T. Denoeux, P. Aknin, Learning from Partially Supervised Data Using Mixture Models and Belief Functions, *Pattern Recognition* 42 (2009) 334–348.
- [15] R. Tibshirani, G. Walther, T. Hastie, Estimating the Number of Clusters in a Data Set via the Gap Statistic, *Journal of the Royal Statistical Society, Series B* 63 (2001) 411–423.
- [16] J. Ma, T. Wang, A Cost-Function Approach to Rival Penalized Competitive Learning (RPCL), *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics* 36 (2006) 722–737.
- [17] N. Bouguila, D. Ziou, High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (2007) 1716–1731.
- [18] C. E. Rasmussen, The Infinite Gaussian Mixture Model, in: *Advances in Neural Information Processing Systems (NIPS)*, MIT Press, 2000, pp. 554–560.
- [19] R. M. Korwar, M. Hollander, Contributions To The Theory Of Dirichlet Processes, *Ann. Probab.* 1 (1973) 705–711.
- [20] M. D. Escobar, Estimating Normal Means with a Dirichlet Process Prior, *Journal of the American Statistical Association* 89 (1994) 268–277.
- [21] T. S. Ferguson, Bayesian Density Estimation by Mixtures of Normal Distributions, *Recent Advances in Statistics*, H. Rizvi and J. Rustagi, Eds. 24 (1983) 287–302.
- [22] C. Robert, G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, 1999.
- [23] D. B. Dunson, Bayesian Semiparametric Isotonic Regression for Count Data, *Journal of the American Statistical Association* 100 (2005) 618–627.
- [24] N. Bouguila, D. Ziou, A Dirichlet Process Mixture of Generalized Dirichlet Distributions for Proportional Data Modeling, *IEEE Transactions on Neural Networks* 21 (2010) 107–122.
- [25] R. M. Neal, Markov Chain Sampling Methods For Dirichlet Process Mixture Models, *Journal of Computational and Graphical Statistics* 9 (2000) 249–265.
- [26] Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei, Hierarchical Dirichlet Processes, *Journal of the American Statistical Association* 101 (2004) 705–711.
- [27] C. E. Antoniak, Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems, *Annals of Statistics* 2 (1974) 1152–1174.
- [28] S. N. MacEachern, P. Müller, Estimating Mixture of Dirichlet Process Models, *J. Comput. Graph. Statist* 7 (1998) 227–238.
- [29] C. E. Rasmussen, Z. Ghahramani, Occam’s Razor, in: *Advances in Neural Information Processing Systems (NIPS)*, MIT Press, 2000, pp. 294–300.
- [30] D. Fragoudis, D. Meretakis, S. Likothanassis, Integrating Feature and Instance Selection for Text Classification, in: *Proc. of the 8th ACM SIGKDD conference on Knowledge discovery and data mining (KDD)*, ACM, 2002, pp. 501–506.
- [31] D. P. Foster, R. A. Stine, Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy, *Journal of the American Statistical Association* 99 (2004) 303–313.
- [32] Y. Wu, A. Zhang, Feature Selection for Classifying High-Dimensional Numerical Data, in: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 2004, pp. 251–258.
- [33] Y. Kim, W. N. Street, F. Menczer, Feature Selection in Unsupervised Learning via Evolutionary Search, in: *Proc. of the 6th ACM SIGKDD conference on Knowledge discovery and data mining (KDD)*, ACM, 2000, pp. 365–369.
- [34] J. M. Pena, J. A. Lozano, P. Larranaga, I. Inza, Dimensionality Reduction in Unsupervised Learning of Conditional Gaussian Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001) 590–603.
- [35] V. Roth, T. Lange, Bayesian Class Discovery in Microarray Datasets, *IEEE Transactions on Biomedical Engineering* 51 (2004) 707–718.

- [36] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, M. W. Mahoney, Feature Selection Methods for Text Classification, in: Proc. of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2007, pp. 230–239.
- [37] J. Zhou, D. Foster, R. Stine, L. Ungar, Streaming Feature Selection Using Alpha-Investing, in: Proc. of the 11th ACM SIGKDD conference on Knowledge discovery and data mining (KDD), ACM, 2005, pp. 384–393.
- [38] C. Bouveyron, S. Girard, C. Schmid, High-Dimensional Data Clustering, *Computational Statistics and Data Analysis* 52 (2007) 502–519.
- [39] D. B. Dunson, A. H. Herring, S. M. Engel, Bayesian Selection and Clustering of Polymorphisms in Functionally Related Genes, *Journal of the American Statistical Association* 103 (2005) 534–546.
- [40] M. Bressan, J. Vitrà, On the Selection and Classification of Independent Features, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (2003) 1312–1317.
- [41] M. H. C. Law, M. A. T. Figueiredo, A. K. Jain, Simultaneous Feature Selection and Clustering Using Mixture Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (2004) 1154–1166.
- [42] C. Constantinopoulos, M. K. Titsias, A. Likas, Bayesian Feature and Model Selection for Gaussian Mixture Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (2006) 1013–1018.
- [43] S. Wang, J. Zhu, Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data, *Biometrics* 64 (2008) 440–448.
- [44] N. Bouguila, D. Ziou, A Countably Infinite Mixture Model for Clustering and Feature Selection, *Knowledge and Information Systems* 33 (2012) 351–370.
- [45] M. Jordan, Z. Ghahramani, T. S. Jaakola, L. K. Saul, An Introduction to Variational Methods for Graphical Models, *Machine Learning* 37 (1999) 183–233.
- [46] Z. Ghahramani, M. J. Beal, Propagation Algorithms for Variational Bayesian Learning, in: *Advances in Neural Information Processing Systems (NIPS)*, MIT Press, 2000, pp. 507–513.
- [47] D. M. Blei, M. I. Jordan, Variational Inference For Dirichlet Process Mixtures, *Bayesian Analysis* 1 (2005) 121–144.
- [48] J. Sethuraman, A Constructive Definition of Dirichlet Priors, *Statistica Sinica* 4 (1994) 639–650.
- [49] W. D. Penny, S. J. Roberts, Variational Bayes for Non-Gaussian Autoregressive Models, in: Proc. of the IEEE Signal Processing Society Workshop On Neural Networks for Signal Processing (NNSP), IEEE Signal Processing Society, 2000, pp. 135–144.
- [50] H. Attias, A Variational Bayes Framework for Graphical Models, in: *Advances in Neural Information Processing Systems (NIPS)*, MIT Press, 1999, pp. 209–215.
- [51] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- [52] N. Bouguila, D. Ziou, A Hybrid Sem Algorithm for High-Dimensional Unsupervised Learning Using a Finite Generalized Dirichlet Mixture, *IEEE Transactions on Image Processing* 15 (2006) 2657–2668.
- [53] H. Ishwaran, L. F. James, Gibbs Sampling Methods for Stick-Breaking Priors, *Journal of the American Statistical Association* 96 (2001) 161–173.
- [54] H. Ishwaran, L. F. James, Some Further Developments for Stick-Breaking Priors: Finite and Infinite Clustering and Classification, *Shankhya: The Indian Journal of Statistics* 65 (2003) 577–592.
- [55] J. M. Dickey, Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses, *Journal of the American Statistical Association* 78 (1983) 628–637.
- [56] Z. Ma, A. Leijon, Bayesian Estimation of Beta Mixture Models with Variational Inference, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (2011) 2160–2173.
- [57] M. A. R. Leisink, H. J. Kappen, General Lower Bounds based on Computer Generated Higher Order Expansions, in: Proc. of the Conference in Uncertainty in Artificial Intelligence (UAI), Morgan Kaufmann, 2002, pp. 293–300.
- [58] C. M. Bishop, M. E. Tipping, Variational Relevance Vector Machines, in: Proc. of the Conference in Uncertainty in Artificial Intelligence (UAI), Morgan Kaufmann, 2000, pp. 46–53.
- [59] Z. Su, H. Zhang, S. Li, S. Ma, Relevance Feedback in Content-Based Image Retrieval: Bayesian Framework, Features Subspaces, and Progressive Learning, *IEEE Transactions on Image Processing* 12 (2003) 924–937.
- [60] S. Boutemedjet, D. Ziou, N. Bouguila, A Graphical Model for Content Based Image Suggestion and Feature Selection, in: J. N. Kok, J. Koronacki, R. L. de Mántaras, S. Matwin, D. Mladenic, A. Skowron (Eds.), *PKDD*, volume 4702 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 30–41.
- [61] J. Matas, J. Burianek, J. Kittler, Object Recognition using the Invariant Pixel-Set Signature, in: Proc. of BMVC, British Machine Vision Association, 2000, pp. 606–615.
- [62] V. Lepetit, P. Fua, Keypoint Recognition Using Randomized Trees, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (2006) 1465–1479.
- [63] G. Ssurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, Visual Categorization with Bags of Keypoints, in: *Workshop on Statistical Learning in Computer Vision*, 8th European Conference on Computer Vision (ECCV), Springer, 2004.
- [64] A. Bosch, A. Zisserman, X. Munoz, Scene Classification Via pLSA, in: Proc. of 9th European Conference on Computer Vision (ECCV), Springer, 2006, pp. 517–530.
- [65] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, A. Zisserman, Using Multiple Segmentations to Discover Objects and their Extent in Image Collections, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2006, pp. 1605–1614.
- [66] T. Hofmann, Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning* 42 (2001) 177–196.
- [67] L. Cao, L. Fei-Fei, Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes, in: Proc. of the IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, 2007, pp. 1–8.
- [68] J. Shi, J. Malik, Normalized Cuts and Image Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000) 888–905.
- [69] K. Mikolajczyk, B. Leibe, B. Schiele, Local Features for Object Class Recognition, in: Proc. of the IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, 2005, pp. 1792–1799 Vol. 2.

- [70] K. Mikolajczyk, C. Schmid, Scale and Affine Invariant Interest Point Detectors, *International Journal of Computer Vision* 60 (2004) 63–86.
- [71] R. Fergus, P. Perona, A. Zisserman, Object Class Recognition By Unsupervised Scale-Invariant Learning, in: *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 2003, pp. 264–271.
- [72] L.-J. Li, L. Fei-Fei, What, Where And Who? Classifying Events by Scene and Object Recognition, in: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 2007, pp. 1–8.
- [73] A. Bosch, A. Zisserman, X. Muoz, Scene Classification Using a Hybrid Generative/Discriminative Approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (2008) 712–727.
- [74] R. Zhao, W. I. Grosky, From Features to Semantics: Some Preliminary Results, in: *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, IEEE Computer Society, 2000, pp. 679–682.
- [75] M. R. Naphade, T. S. Huang, A Probabilistic Framework for Semantic Video Indexing, Filtering, and Retrieval, *IEEE Transactions on Multimedia* 3 (2001) 141–151.
- [76] E. Chang, CBSA: Content-Based Soft Annotation for Multinomial Image Retrieval using Bayes Point Machines, *IEEE Transactions on Circuit and Systems for Video Technology* 13 (2003) 26–38.
- [77] J. Luo, A. E. Savakis, A. Singhal, A Bayesian Network-Based Framework for Semantic Image Understanding, *Pattern Recognition* 38 (2005) 919–934.
- [78] J. Fan, Y. Gao, H. Luo, G. Xu, Statistical Modeling and Conceptualization of Natural Images, *Pattern Recognition* 38 (2005) 865–885.
- [79] P. H. Gosselin, M. Cord, Feature-Based Approach to Semi-Supervised Similarity Learning, *Pattern Recognition* 39 (2006) 1839–1851.
- [80] N. Hervé, N. Boujemaa, Image Annotation: Which Approach for Realistic Databases?, in: *Proc. of the 6th ACM International Conference on Image and Video Retrieval (CIVR)*, ACM, 2007, pp. 170–177.
- [81] K. Barnard, D. Forsyth, Learning the Semantics of Words and Pictures, in: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 2001, pp. 408–415.
- [82] F. Monay, D. Gatica-Perez, On Image Auto-Annotation with Latent Space Models, in: *Proc. of the eleventh ACM international conference on Multimedia (MM)*, ACM, 2003, pp. 275–278.
- [83] E. P. Xing, R. Yan, A. G. Hauptmann, Mining Associated Text and Images with Dual-Wing Harmoniums, in: *Proc. of the Conference in Uncertainty in Artificial Intelligence (UAI)*, AUAI Press, 2005, pp. 633–641.
- [84] J. Li, A Mutual Semantic Endorsement Approach to Image Retrieval and Context Provision, in: *Proc. of the 7th ACM SIGMM international workshop on Multimedia information retrieval (MIR)*, ACM, 2005, pp. 173–182.
- [85] R. Datta, W. Ge, J. Li, J. Z. Wang, Toward Bridging The Annotation-Retrieval Gap in Image Search by A Generative Modeling Approach, in: *Proc. of the 14th annual ACM international conference on Multimedia (MM)*, ACM, 2006, pp. 977–986.
- [86] Y. Ke, R. Sukthankar, Pca-Sift: A More Distinctive Representation for Local Image Descriptors, in: *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 2004, pp. 506–513.
- [87] G. A. Miller, WordNet: A Lexical Database for English, *Communications of the ACM* 38 (1995) 39–41.
- [88] C. Leacock, M. Chodorow, WordNet: An Electronic Lexical Database, In C. Fellbaum (Ed.), MIT Press, 1998.
- [89] B. Russell, A. Torralba, K. Murphy, W. Freeman, LabelMe: A Database and Web-Based Tool for Image Annotation, *International Journal of Computer Vision* 77 (2008) 157–173.