Appl Intell
DOI 10.1007/s10489-015-0714-6

Variational Bayesian inference for infinite generalized inverted Dirichlet mixtures with feature selection and its application to clustering

Taoufik Bdiri[1] · **Nizar Bouguila**[2] · **Djemel Ziou**[3]

Abstract We developed a variational Bayesian learning framework for the infinite generalized Dirichlet mixture model (i.e. a weighted mixture of Dirichlet process priors based on the generalized inverted Dirichlet distribution) that has proven its capability to model complex multidimensional data. We also integrate a "feature selection" approach to highlight the features that are most informative in order to construct an appropriate model in terms of clustering accuracy. Experiments on synthetic data as well as real data generated from visual scenes and handwritten digits datasets illustrate and validate the proposed approach.

Keywords Data clustering · Mixture models · Variational Bayesian inference · Generalized inverted Dirichlet · Inverted beta · Model selection · Visual scenes · Handwritten digits

✉ Nizar Bouguila
nizar.bouguila@concordia.ca

Taoufik Bdiri
t_bdiri@live.concordia.ca

Djemel Ziou
djemel.ziou@usherbrooke.ca

[1] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, H3G 1T7, Canada

[2] The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, H3G 1T7 Canada

[3] DI, Faculté des Sciences, Université de Sherbrooke, Sherbrooke, QC, J1K 2R1, Canada

1 Introduction

Data analysis is crucial for any business survival as it represents the basis of data-driven decisions that improve and maintain the quality of a given provided service or production. Indeed, with the increase of calculations power, data analysis is now present in many sectors including but not limited to health care, IT services, drug manufactures and finance. Amongst others, data analysis is strongly applied in modern applications that cover computer vision and object recognition. As a matter of fact, data generation has become a daily routine in many industries and entertainment services, e.g. 300 hours of video are uploaded, every minute, to YouTube, a service that has more than 1 billion users (https://www.youtube.com/yt/press/statistics.html). Clustering is an important exploratory technique of data analysis where observations are assigned to different groups that contain similar observations [1]. Many clustering techniques have been explored by researchers [2] especially model-based clustering that has been extensively adopted in order to perform data analysis in many disciplines such as bioinformatics [3–5], computer vision [6–8], recommending systems [9–11] and social networks [12–14]. Indeed, model based approaches are a powerful tool that serve to infer knowledge and abstract the complexity of a huge amount of information. They are a well formed mathematical representation that assumes that data are generated by a set of probabilistic components representing subpopulations and that every observation belongs to one of them. Such an assumption establishes a clustering process based on posteriors probabilities. Yet, the adoption of model based approaches is still under extensive research, and the task of their deployment faces many challenges, such as the choice of the appropriate distribution to model data, how to estimate the parameters of the mixture and how to select the

Published online: 06 October 2015

model that is appropriate to better describe data in terms of number of components.

We have shown in our previous works [15–19], that the conventional use of gaussian mixture model (GMM) is not always appropriate to model data when the partitions are not Gaussian, and that the inverted Dirichlet (ID) mixture model and generalized inverted Dirichlet (GID) mixture model often outperform the GMM is terms of clustering accuracy. In [18], we have proposed a new methodology in order to update a given model when new data arrive on line. The approach proposes to establish a first model of the arriving data, then compare its components with the existing components in the model, and decide whether 1) to create new components in the final model, or 2) update the existing components of the old model, or, 3) do both. We have also introduced a user perception parameter that can help the system to have a data representation that uses a hierarchical model which groups sub-clusters in order to form a parent cluster that does not necessarily have similar data in the system feature space but have the same meaning in the users semantic [17, 18]. Still the challenging point in the proposed model is to create the model of new data that are arriving. Previously, in [18], we have considered five criteria in order to establish model selection, namely the minimum message length (MML) [20], Akaike information criterion (AIC) [21], minimum description length (MDL) [22], mixture MDL (MMDL) [23], and LEC [24]. Yet, these approaches are demanding in terms of computing cost as we have to establish N complete estimations in order to select a model among N models. Also the use of Newton-Raphson technique within the expectation maximization (EM) framework [25] is not always performant, as 1) it is not guaranteed to converge in general, 2) is prone to finding poor solutions in the case of multi-modal distributions [26] and 3) is dependent on initialization [24, 27]. In order to mitigate these problems many researches considered pure Bayesian approaches, where the parameters are considered to be random variables and then follow probability distributions called priors that describe our knowledge before using data [28], e.g. Bayesian frameworks using Markov chain Monte Carlo (MCMC) have been proposed in [29, 30], and a non parametric Bayesian approach based on Dirichlet processes is proposed in [31]. The main concern about fully Bayesian approaches is that they are computationally costly and their convergence is difficult to assess [32, 33]. In order to overcome those problems, several variational Bayesian techniques has been proposed [34–37]. Indeed, the variational Bayesian inference consists of estimating a lower bound for the likelihood of observed data with a marginalization performed over unobserved variables and it constitutes a better alternative to MCMC. The other interesting aspect when performing clustering is to establish feature selection. Indeed, it has been shown that not all the features have the same contribution in the clustering process such as in [38] where the generalized Dirichlet (GD) mixture has been used and factorized into a set of Beta distributions giving good results for proportional data [39] clustering. Still, the Beta distribution support is defined in [0 1], so it is not always an appropriate choice to represent positive data. The GID mixture that can be factorized into a set of Beta prime (inverted Beta) distributions whose support is $]0\ \infty[$, is still a more adequate choice to represent positive data [40]. We propose in this paper to build a variational Bayesian framework of GID mixture with feature selection, and investigate its modeling capabilities using synthetic and real data.

The rest of this paper is organized as follows; in Section 2 we introduce the statistical model that is based on an infinite GID mixture model with feature selection.Then, we propose a variational Bayesian inference that estimates the parameters of the proposed model in Section 3. In Section 4 we investigate the performance of our algorithm on synthetic and real data and we conclude in Section 5.

2 The statistical model

2.1 Finite generalized inverted Dirichlet mixture model

Let us consider a set $\mathcal{Y}$ of N D-dimensional vectors, such that $\mathcal{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N)$. Let M denote the number of different components forming a flat mixture model [24] at the system level. We assume that $\mathcal{Y}$ is controlled by a mixture of GID distributions and the vectors follow a common probability density function $p(\mathbf{Y}_i|\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ such that

$$p(\mathbf{Y}_i|\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{M} \pi_j \prod_{d=1}^{D} \frac{\Gamma(\alpha_{jd} + \beta_{jd})}{\Gamma(\alpha_{jd})\Gamma(\beta_{jd})} \frac{Y_{id}^{\alpha_{jd}-1}}{(1 + \sum_{l=1}^{d} Y_{il})^{\gamma_{jd}}} \tag{1}$$

where $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_M\}$, with $\boldsymbol{\alpha}_j = \{\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jD}\}$, $j = 1, .., M$, and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_M\}$, with $\boldsymbol{\beta}_j = \{\beta_{j1}, \beta_{j2}, ..., \beta_{jD}\}$, $j = 1, .., M$. $\boldsymbol{\pi} = \{\pi_1, \pi_2, ..., \pi_M\}$ are the mixing weights, such that $\sum_{j=1}^{M} \pi_j = 1$. We define γ_{jd} such that $\gamma_{jd} = \beta_{jd} + \alpha_{jd} - \beta_{j(d+1)}$. The GID posterior probability can be factorized as follows (see Appendix A)

$$p(j|\mathbf{Y}_i, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \pi_j \prod_{l=1}^{D} p_{iBeta}(X_{il}|\alpha_{jl}, \beta_{jl}) \tag{2}$$

where we have set $X_{i1} = Y_{i1}$ and $X_{il} = \frac{Y_{il}}{1+\sum_{k=1}^{l-1} Y_{ik}}$ for $l > 1$. $p_{iBeta}(X_{il}|\alpha_{jl}, \beta_{jl})$ is an inverted Beta distribution with parameters α_{jl} and β_{jl}

$$p_{iBeta}(X_{il}|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) = \frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} X_{il}^{\alpha_{jl}-1}(1+X_{il})^{-(\alpha_{jl}+\beta_{jl})} \tag{3}$$

Thus, the clustering structure underlying $\mathcal{Y}$ is the same as the one underlying $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N\}$, where $\mathbf{X}_i = \{X_{i1}, X_{i2}, ..., X_{iD}\}$, $i = 1, ..., N$, governed by the following mixture model with conditionally independent features

$$p(\mathcal{X}|\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})) = \prod_{i=1}^{N}\left(\sum_{j=1}^{M}\pi_j \prod_{l=1}^{D} p_{iBeta}(X_{il}|\alpha_{jl}, \beta_{jl})\right) \tag{4}$$

The estimation of the parameters in (1) is then equivalent to the estimation of the parameters in (4).

2.2 Infinite generalized inverted Dirichlet mixture model

We construct an infinite GID model using the a Dirichlet Process with a stick-breaking representation [41] such that

$$p(X_i|\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})) = \sum_{j=1}^{\infty}\pi_j \prod_{l=1}^{D} p_{iBeta}(X_{il}|\alpha_{jl}, \beta_{jl}) \tag{5}$$

with $\pi_j = \lambda_j \prod_{s=1}^{j-1}(1-\lambda_s)$ and $\lambda_j \sim Beta(1, \psi)$ where ψ is a real number. Let $\mathcal{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N\}$ denote the missing group indicator, where $\mathbf{Z}_n = (z_{n1}, z_{n2}, ...)$ is the label of $\mathbf{X}_n$, such that $z_{nj} \in \{0, 1\}$, $\sum_{j=1}^{\infty} z_{nj} = 1$ and z_{nj} is equal to one if $\mathbf{X}_n$ belongs to class j and zero, otherwise. Then, the distribution of $\mathcal{X}$ given the class label Z is

$$p(\mathcal{X}|Z, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{N}\prod_{j=1}^{\infty}\left(\prod_{l=1}^{D} p_{iBeta}(X_{il}|\alpha_{jl}, \beta_{jl})\right)^{z_{ij}} \tag{6}$$

2.3 Infinite generalized inverted Dirichlet mixture model with feature selection

Feature selection is a fundamental aspect of machine learning especially with the growth of data dimensionality [42]. Indeed, when data is multidimensional some of the features could be noisy to the clustering process and deteriorate its performance. A feature is irrelevant when it does not have a discriminatory effect on the clusters. Mathematically speaking, if we consider M clusters, and $\forall n, m \in \{1, ..., M\}$, $KL\left(p_{iBeta}(.|\alpha_{nl}, \beta_{nl}), p_{iBeta}(.|\alpha_{ml}, \beta_{ml})\right) \simeq 0$ where KL is the Kullback-Leibler divergence, then $\{\alpha_{jl}, \beta_{jl}\} = \{\sigma_l, \tau_l\}$, $\forall j$, such that $\{\sigma_l, \tau_l\}$ are the parameters of a common inverted Beta distribution independent from the class labels. The works in [43, 44] have considered a common univariate distribution to model the irrelevant features, while the works in [38] and [45] have respectively considered finite and infinite mixtures of overlapped distributions to be common to all clusters than a single univariate distribution. We adopt the infinite mixture model to represent the irrelevant features, and therefore we approximate each features as follows

$$p(X_{il}) \simeq \left(iBeta(X_{il}|\alpha_{jl}, \beta_{jl})\right)^{\phi_{il}} \times \left(\prod_{k=1}^{\infty}\left(iBeta(X_{il}|\sigma_{kl}, \tau_{kl})\right)^{W_{ikl}}\right)^{1-\phi_{il}} \tag{7}$$

where $W_{ikl} \in \{1, 0\}$ such that W_{ikl} is equal to one if X_{il} is generated from the k^{th} components of the infinite Beta mixture representing the irrelevant features. $\phi_{il} \in \{0, 1\}$ and is equal to one when l is a relevant feature and follows an inverted Beta distribution $iBeta(X_{il}|\alpha_{jl}, \beta_{jl})$, otherwise the feature l follows an infinite mixture of inverted Beta distributions such that

$$p\left(X_{il}\right) = \sum_{k=1}^{\infty}\eta_k iBeta\left(X_{il}|\sigma_{kl}, \tau_{kl}\right) \tag{8}$$

where η_k is the mixing probability, and $\{\sigma_{kl}, \tau_{kl}\}$ are the parameters of the inverted Beta representing the k^{th} components of the irrelevant feature. The likelihood of $\mathcal{X}$ that follows an infinite GID mixture model with feature selection can be written under the form

$$\begin{aligned} &p(\mathcal{X}|\mathcal{Z}, \mathcal{W}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\tau}) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{\infty}\left[\prod_{l=1}^{D}\left(iBeta\left(X_{il}|\alpha_{jl}, \beta_{jl}\right)\right)^{\phi_{il}} \times \left(\prod_{k=1}^{\infty}\left(iBeta\left(X_{il}|\sigma_{kl}, \tau_{kl}\right)\right)^{W_{ikl}}\right)^{1-\phi_{il}}\right]^{Z_{ij}} \end{aligned} \tag{9}$$

2.4 Prior distributions for the infinite GID mixture with feature selection

The variational Bayesian approach needs the definition of priors for $\mathcal{Z}, \mathcal{W}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}$ and $\boldsymbol{\tau}$. We consider priors that can give us tractable solutions for updating the variational factors. The priors of $\mathcal{Z}$ and $\mathcal{W}$ given the mixing coefficients $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$ can be set as

$$p(\mathcal{Z}|\boldsymbol{\pi}) = \prod_{i=1}^{N}\prod_{j=1}^{\infty}\pi_j^{Z_{ij}} \qquad p(\mathcal{W}|\boldsymbol{\eta}) = \prod_{i=1}^{N}\prod_{k=1}^{\infty}\prod_{l=1}^{D}\eta_k^{W_{ikl}} \tag{10}$$

π_j and η_k can be written according the stick-breaking approach under the following form

$$\pi_j = \lambda_j \prod_{s=1}^{j-1}(1-\lambda_s) \qquad \eta_k = \gamma_k \prod_{s=1}^{k-1}(1-\gamma_s) \tag{11}$$

Using (10) and (11), we can have the following distribution for $\mathcal{Z}$ and $\mathcal{W}$

$$\begin{aligned} p(\mathcal{Z}|\boldsymbol{\lambda}) &= \prod_{i=1}^{N}\prod_{j=1}^{\infty}\left[\lambda_j \prod_{s=1}^{j-1}(1-\lambda_s)\right]^{Z_{ij}} \\ p(\mathcal{W}|\boldsymbol{\gamma}) &= \prod_{i=1}^{N}\prod_{k=1}^{\infty}\prod_{l=1}^{D}\left[\gamma_k \prod_{s=1}^{k-1}(1-\gamma_s)\right]^{W_{ikl}} \end{aligned} \tag{12}$$

The prior distributions of $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are given by the stick-breaking approach as follows

$$\begin{aligned} p(\boldsymbol{\lambda}|\boldsymbol{\psi}) &= \prod_{j=1}^{\infty} Beta(1,\psi_j) = \prod_{j=1}^{\infty}\psi_j(1-\lambda_j)^{\psi_j-1} \\ p(\boldsymbol{\gamma}|\boldsymbol{\phi}) &= \prod_{k=1}^{\infty} Beta(1,\phi_k) = \prod_{j=1}^{\infty}\psi_k(1-\gamma_k)^{\phi_k-1} \end{aligned} \tag{13}$$

where $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ are the hyperparameters to which we can attribute conjugate Gamma priors [46] as follows

$$\begin{aligned} p(\boldsymbol{\psi}) = \mathcal{G}(\boldsymbol{\psi}|\mathbf{a},\mathbf{b}) &= \prod_{j=1}^{\infty}\frac{b_j^{a_j}}{\Gamma(a_j)}\psi^{a_j-1}e^{-b_j\psi_j} \\ p((\phi)) = \mathcal{G}(\boldsymbol{\phi}|\mathbf{c},\mathbf{d}) &= \prod_{k=1}^{\infty}\frac{d_k^{c_k}}{\Gamma(c_k)}\varphi^{c_k-1}e^{-d_k\varphi_k} \end{aligned} \tag{14}$$

The elements of hyperparameters vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ and $\mathbf{d}$ are strictly positive. The prior distribution of the feature indicator variable $\boldsymbol{\phi}$ is defined as

$$p(\boldsymbol{\phi}|\boldsymbol{\epsilon}) = \prod_{i=1}^{N}\prod_{l=1}^{D}\epsilon_{l1}^{\phi_{il}}\epsilon_{l2}^{1-\phi_{il}} \tag{15}$$

where $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, .., \boldsymbol{\epsilon}_D)$ is the vector of feature saliencies such that $\boldsymbol{\epsilon}_l = (\epsilon_{l1}, \epsilon_{l2})$, and $\epsilon_{l1} + \epsilon_{l2} = 1$. $\boldsymbol{\epsilon}$ can be seen as a proportional data so it can be modeled by a Dirichlet distribution such that

$$p(\boldsymbol{\epsilon}) = \prod_{l=1}^{D} Dir(\boldsymbol{\epsilon}_l|\boldsymbol{\xi}) = \prod_{l=1}^{D}\frac{\Gamma(\xi_1+\xi_2)}{\Gamma(\xi_1)\Gamma(\xi_2)}\epsilon_{l_1}^{\xi_1-1}\epsilon_{l_2}^{\xi_2-1} \tag{16}$$

with the hyperparameters $\boldsymbol{\xi} = (\epsilon_{l1}, \epsilon_{l2})$ that are strictly positive. As for the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ we consider the Gamma distribution as a prior distribution for them such that

$$\begin{aligned} p(\boldsymbol{\alpha}) &= \mathcal{G}(\boldsymbol{\alpha}|\mathbf{u},\mathbf{v}) = \prod_{j=1}^{\infty}\prod_{l=1}^{D}\frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})}\alpha_{jl}^{u_{jl}-1}e^{-v_{jl}\alpha_{jl}} \\ p(\boldsymbol{\beta}) &= \mathcal{G}(\boldsymbol{\beta}|\mathbf{p},\mathbf{q}) = \prod_{j=1}^{\infty}\prod_{l=1}^{D}\frac{q_{jl}^{p_{jl}}}{\Gamma(p_{jl})}\beta_{jl}^{p_{jl}-1}e^{-q_{jl}\beta_{jl}} \\ p(\boldsymbol{\sigma}) &= \mathcal{G}(\boldsymbol{\sigma}|\mathbf{g},\mathbf{h}) = \prod_{k=1}^{\infty}\prod_{l=1}^{D}\frac{h_{kl}^{g_{kl}}}{\Gamma(g_{kl})}\sigma_{kl}^{g_{kl}-1}e^{-h_{kl}\sigma_{kl}} \\ p(\boldsymbol{\tau}) &= \mathcal{G}(\boldsymbol{\sigma}|\mathbf{s},\mathbf{t}) = \prod_{k=1}^{\infty}\prod_{l=1}^{D}\frac{t_{kl}^{s_{kl}}}{\Gamma(s_{kl})}\tau_{kl}^{s_{kl}-1}e^{-t_{kl}\tau_{kl}} \end{aligned} \tag{17}$$

where all the elements of hyperparameters vectors $\mathbf{u}, \mathbf{v}, \mathbf{q}, \mathbf{g}, \mathbf{h}, \mathbf{t}$, and $\mathbf{s}$ are strictly positive. To summarize, the set of parameters of unknown variables is $\Theta = \{\mathcal{Z}, \mathcal{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\varphi}, \boldsymbol{\epsilon}\}$, and the joint distribution of all the random variables is given by

$$\begin{aligned} p(\mathcal{X},\Theta) &= p(\mathcal{X}|\mathcal{Z},\mathcal{W},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\epsilon}) \\ &= p(\mathcal{X}|\mathcal{Z},\mathcal{W},\Phi,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\tau})p(\mathcal{Z}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\Psi})p(\boldsymbol{\Psi})p(\mathcal{W}|\boldsymbol{\gamma}) \\ &\times p(\boldsymbol{\gamma}|\boldsymbol{\varphi})p(\boldsymbol{\varphi})p(\boldsymbol{\phi}|\boldsymbol{\epsilon})p(\boldsymbol{\epsilon})p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\sigma})p(\boldsymbol{\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})}\frac{X_{il}^{\alpha_{jl}-1}}{(1+X_{il})^{(\alpha_{jl}+\beta_{jl})}}\right]^{\phi_{il}}\right. \\ &\left.\times\left[\prod_{k=1}^{\infty}\left(\frac{\Gamma(\sigma_{kd}+\tau_{kd})}{\Gamma(\sigma_{kl})\Gamma(\tau_{kl})}\frac{X_{il}^{\sigma_{kl}-1}}{(1+X_{il})^{(\sigma_{kl}+\tau_{kl})}}\right)^{w_{ikl}}\right]^{1-\phi_{il}}\right\}^{Z_{ij}} \\ &\times\prod_{i=1}^{N}\prod_{j=1}^{\infty}\left[\lambda_j\prod_{s=1}^{j-1}(1-\lambda_s)\right]^{Z_{ij}}\times\prod_{j=1}^{\infty}\psi_j(1-\lambda_j)^{\psi_j-1} \\ &\times\prod_{j=1}^{\infty}\frac{b_j^{a_j}}{\Gamma(a_j)}\psi^{a_j-1}e^{-b_j\psi_j}\times\prod_{i=1}^{N}\prod_{k=1}^{\infty}\prod_{l=1}^{D}\left[\gamma_k\prod_{s=1}^{k-1}(1-\gamma_s)\right]^{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^{c_k-1}e^{-d_k\varphi_k}\times\prod_{i=1}^{N}\prod_{l=1}^{D}\epsilon_{l_1}^{\phi_{li}}\epsilon_{l_2}^{1-\phi_{li}} \\ &\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] \\ &\times\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} \tag{18}$$

Figure 1 illustrates the dependencies between all the variables via a graphical model.

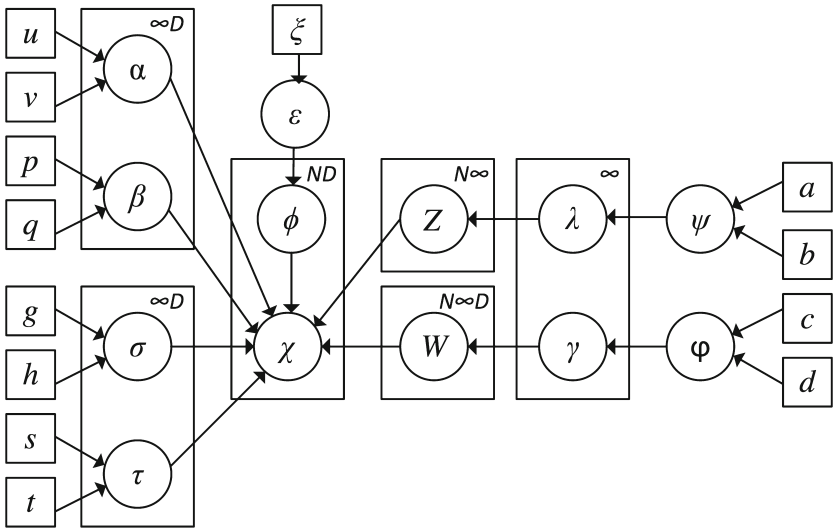


Fig. 1 Graphical model representation of the infinite GID mixture model with feature selection. The random variables are in circles, and the model parameters in squares. The number mentioned in the right upper corner of the plates indicates the number of repetition of the contained random variables. The arcs describe the conditional dependencies between variables

3 Variational inference

In this section we develop a variational inference framework for the parameters estimation of the infinite GID mixture with feature selection. The aim of variational inference is to determine a distribution $Q(\Theta)$ that approximates the true posterior distribution $p(\Theta|\mathcal{X})$. The Kullback-Leibler (KL) divergence between $Q(\Theta)$ and $p(\Theta|\mathcal{X})$ is given by

$$KL(Q||P) = -\int Q(\Theta)\ln\left(\frac{p(\Theta|\mathcal{X})}{Q(\Theta)}\right)d\Theta = \ln p(\mathcal{X}) - \mathcal{L}(Q) \tag{19}$$

with

$$\mathcal{L}(Q) = \int Q(\Theta)\ln\left(\frac{p(\Theta,\mathcal{X})}{Q(\Theta)}\right)d\Theta \tag{20}$$

since $KL(Q||P) \geq 0$ and equal to zero when $Q(\Theta) = p(\Theta|\mathcal{X})$, we can conclude from (19) that $\mathcal{L}(Q) \leq \ln p(\mathcal{X})$, which means that $\mathcal{L}(Q)$ can be considered as a lower bound to $\ln p(\mathcal{X})$. However in practice, the true posterior distribution is computationally intractable and cannot be directly used for variational inference. Thus, we have to consider a restricted family of $Q(\Theta)$ that can be computed [47]. Indeed we factorize $Q(\Theta)$ into disjoint tractable distributions such that $Q(\Theta) = \prod_i Q_i(\Theta_i)$, which is known as the mean field theory [48]. The maximization of $\mathcal{L}(Q)$ is established through a variational optimization with respect to each of the factor distributions $Q_i(\Theta_i)$. If we consider a specific Q_s, the variational approximation consists of keeping $\{\Theta_i\}_{i\neq s}$ fixed and maximize $\mathcal{L}(Q)$ with respect to all possible forms for the distribution $Q_s(\Theta_s)$. The optimal solution for $Q_s(\Theta_s)$ is given by [49]

$$\ln Q_s(\Theta_s) = \langle \ln p(\mathcal{X},\Theta)\rangle_{j\neq s} + const \tag{21}$$

where $\langle . \rangle_{j\neq s}$ denotes an expectation with respect to all the distributions $Q_i(\Theta_i)$ except for $i = s$, such that

$$\langle \ln p(\mathcal{X},\Theta)\rangle_{j\neq s} = \int \ln p(\mathcal{X},\Theta)\prod_{i\neq s} Q_i(\Theta_i)d\Theta_i \tag{22}$$

and the normalized solution is given by

$$Q_s(\Theta_s) = \frac{\exp\langle \ln p(\mathcal{X},\Theta)\rangle_{j\neq s}}{\exp\int \langle \ln p(\mathcal{X},\Theta)\rangle_{j\neq s} d\Theta} \tag{23}$$

As proposed by [49], we initialize $Q_s(\Theta_s)$ appropriately, and then we cycle through the factors and replace each in turn with a revised estimate given by (21) and evaluated using the current estimates for all the other factors. A convergence is guaranteed because the bound is convex with respect to each of the factors $Q_i(\Theta_i)$ [49, 50]. In order to exploit the bound, we should consider the truncation of the stick-breaking representation as proposed by [46] to establish a variational inference for DP mixtures. The truncation of the stick-breaking representation was also proposed by [51] in the context of sampling-based inference for an approximation to the DP mixture model. This is done by fixing a value M, such that $\lambda_M = 1$, and $\pi_j = 0$ when $j > M$, which leads to $\sum_{j=1}^{M}\pi_j = 1$. We apply a similar truncation to the infinite inverted Beta mixture representing the irrelevant features by fixing a value K, such that $\gamma_K = 1$, and $\eta_k = 0$ when $k > K$, which leads to $\sum_{k=1}^{K}\eta_k = 1$. Note that the model still a full Dirichlet process and is not truncated, only the variational distribution is truncated. The truncation levels M and K are variational parameters that can be freely set, and they are not a part of the prior model specification [49]. Thus M and K can be optimized during

the learning process. The factorization of $Q(\Theta)$ can be written under the following form

$$
\begin{aligned}
Q(\Theta) = & \left[\prod_{i=1}^{N}\prod_{j=1}^{M} Q(Z_{ij})\right]\left[\prod_{j=1}^{M} Q(\lambda_j)Q(\Psi_j)\right] \\
& \left[\prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{l=1}^{D} Q(W_{ikl})\right] \\
& \times\left[\prod_{k=1}^{K} Q(\gamma_k)Q(\varphi_k)\right]\left[\prod_{i=1}^{N}\prod_{l=1}^{D} Q(\phi_{il})\right]\left[\prod_{l=1}^{D} Q(\boldsymbol{\epsilon}_l)\right] \\
& \times\left[\prod_{j=1}^{M}\prod_{l=1}^{D} Q(\alpha_{jl})Q(\beta_{jl})\right]\left[\prod_{k=1}^{K}\prod_{l=1}^{D} Q(\sigma_{kl})Q(\tau_{kl})\right]
\end{aligned} \tag{24}
$$

The optimal solutions are the following, (details in Appendix B)

$$
\begin{aligned}
Q(Z) &= \textstyle\prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}} & Q(\boldsymbol{\lambda}) &= \prod_{j=1}^{M} Beta(\lambda_j|\theta_j, v_j) \\
Q(\boldsymbol{\psi}) &= \textstyle\prod_{j=1}^{M} \mathcal{G}(\psi_j|a_j^*, b_j^*) & Q(\boldsymbol{\alpha}) &= \prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\alpha_{jl}|u_{jl}^*, v_{jl}^*) \\
Q(\boldsymbol{\beta}) &= \textstyle\prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\beta_{jl}|p_{jl}^*, q_{jl}^*) & Q(W) &= \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{l=1}^{D} m_{ikl}^{W_{ikl}} \\
Q(\boldsymbol{\gamma}) &= \textstyle\prod_{k=1}^{K} Beta(\gamma_k|\rho_k, \varpi_k) & Q(\boldsymbol{\varphi}) &= \prod_{k=1}^{K} \mathcal{G}(\varphi_k|c_k^*, d_k^*) \\
Q(\boldsymbol{\phi}) &= \textstyle\prod_{i=1}^{N}\prod_{l=1}^{D} f_{il}^{\phi_{il}}(1-f_{il})^{(1-\phi_{il})} & Q(\boldsymbol{\epsilon}) &= \prod_{l=1}^{D} Dir(\epsilon_l|\boldsymbol{\xi}^*) \\
Q(\boldsymbol{\sigma}) &= \textstyle\prod_{k=1}^{K}\prod_{l=1}^{D} \mathcal{G}(\sigma_{kl}|g_{kl}^*, h_{kl}^*) & Q(\boldsymbol{\tau}) &= \prod_{k=1}^{K}\prod_{l=1}^{D} \mathcal{G}(\tau_{kl}|s_{kl}^*, t_{kl}^*)
\end{aligned} \tag{25}
$$

with

$$
\begin{aligned}
r_{ij} &= \frac{\tilde{r}_{ij}}{\sum_{j=1}^{M}\tilde{r}_{ij}}, \quad m_{ikl} = \frac{\tilde{m}_{ikl}}{\sum_{k=1}^{K}\tilde{m}_{ikl}}, \\
f_{il} &= \frac{f_{il}^{(\phi_{il})}}{f_{il}^{(\phi_{il})} + f_{il}^{(1-\phi_{il})}}
\end{aligned} \tag{26}
$$

$$
\begin{aligned}
\tilde{r}_{ij} = \exp\Bigg\{ & \sum_{l=1}^{D}\Big(\langle\phi_{il}\rangle\left[\tilde{\mathcal{R}}_{jl} + (\bar{\alpha}_{jl}-1)\ln X_{il} - (\bar{\alpha}_{jl}+\bar{\beta}_{jl})\ln(1+X_{il})\right] \\
& + \langle 1-\phi_{il}\rangle\sum_{k=1}^{K}\langle w_{ikl}\rangle\left[\tilde{\mathcal{F}}_{kl} + (\bar{\sigma}_{kl}-1)\ln X_{il} - (\bar{\sigma}_{kl}+\bar{\tau}_{kl})\ln(1+X_{il})\right]\Big) \\
& + \langle\ln\lambda_j\rangle + \sum_{s=1}^{j-1}\langle\ln(1-\lambda_s)\rangle\Bigg\}
\end{aligned} \tag{27}
$$

$$
\begin{aligned}
\tilde{m}_{ikl} = \exp\Bigg\{ & \langle 1-\phi_{il}\rangle\left(\mathcal{F}_{kl} + (\bar{\sigma}_{kl}-1)\ln X_{il}\right. \\
& \left. - (\bar{\sigma}_{kl}+\bar{\tau}_{kl})\ln(1+X_{il})\right) + \langle\ln\gamma_k\rangle + \sum_{s=1}^{k-1}\langle\ln(1-\gamma_s)\rangle\Bigg\}
\end{aligned} \tag{28}
$$

$$
\begin{aligned}
f_{il}^{(\phi_{il})} = \exp\Bigg\{ & \langle\ln\epsilon_{l1}\rangle + \sum_{j=1}^{M}\langle Z_{ij}\rangle\left[\mathcal{R}_{jl} + (\bar{\alpha}_{jl}-1)\ln X_{il}\right. \\
& \left. - (\bar{\alpha}_{jl}+\bar{\beta}_{jl})\ln(1+X_{il})\right]\Bigg\}
\end{aligned} \tag{29}
$$

$$
\begin{aligned}
f_{il}^{(1-\phi_{il})} = \exp\Bigg\{ & \langle\ln\epsilon_{l2}\rangle + \Bigg\{\sum_{k=1}^{K}\langle W_{ikl}\rangle\left[\mathcal{F}_{kl} + (\bar{\sigma}_{kl}-1)\ln X_{il}\right. \\
& \left. - (\bar{\sigma}_{kl}+\bar{\tau}_{kl})\ln(1+X_{il})\right]\Bigg\}\Bigg\}
\end{aligned} \tag{30}
$$

$$
\begin{aligned}
\tilde{\mathcal{R}} = & \ln\frac{\Gamma(\bar{\alpha}+\bar{\beta})}{\Gamma(\bar{\alpha})\Gamma(\bar{\beta})} + \bar{\alpha}[\psi(\bar{\alpha}+\bar{\beta}) - \psi(\bar{\alpha})](\langle\ln\alpha\rangle - \ln\bar{\alpha}) \\
& + \bar{\beta}[\psi(\bar{\beta}+\bar{\alpha}) - \psi(\bar{\beta})](\langle\ln\beta\rangle - \ln\bar{\beta}) \\
& + 0.5\bar{\alpha}^2[\psi'(\bar{\alpha}+\bar{\beta}) - \psi'(\bar{\alpha})]\langle(\ln\alpha - \ln\bar{\alpha})^2\rangle \\
& + 0.5\bar{\beta}^2[\psi'(\bar{\beta}+\bar{\alpha}) - \psi'(\bar{\beta})]\langle(\ln\beta - \ln\bar{\beta})^2\rangle \\
& + \bar{\alpha}\bar{\beta}\psi'(\bar{\alpha}+\bar{\beta})(\langle\ln\alpha\rangle - \ln\bar{\alpha})(\langle\ln\beta\rangle - \ln\bar{\beta})
\end{aligned} \tag{31}
$$

$$
\begin{aligned}
\tilde{\mathcal{F}} = & \ln\frac{\Gamma(\bar{\sigma}+\bar{\tau})}{\Gamma(\bar{\sigma})\Gamma(\bar{\tau})} + \bar{\sigma}[\psi(\bar{\sigma}+\bar{\tau}) - \psi(\bar{\sigma})](\langle\ln\sigma\rangle - \ln\bar{\sigma}) \\
& + \bar{\tau}[\psi(\bar{\tau}+\bar{\sigma}) - \psi(\bar{\tau})](\langle\ln\tau\rangle - \ln\bar{\tau}) \\
& + 0.5\bar{\sigma}^2[\psi'(\bar{\sigma}+\bar{\tau}) - \psi'(\bar{\sigma})]\langle(\ln\sigma - \ln\bar{\sigma})^2\rangle \\
& + 0.5\bar{\tau}^2[\psi'(\bar{\tau}+\bar{\sigma}) - \psi'(\bar{\tau})]\langle(\ln\tau - \ln\bar{\tau})^2\rangle \\
& + \bar{\sigma}\bar{\tau}\psi'(\bar{\sigma}+\bar{\tau})(\langle\ln\sigma\rangle - \ln\bar{\sigma})(\langle\ln\tau\rangle - \ln\bar{\tau})
\end{aligned} \tag{32}
$$

where $\psi(.)$ is the digamma function that is defined as $\psi(\alpha) = \frac{d\ln\Gamma(\alpha)}{d\alpha}$.

$$
\theta_j = 1 + \sum_{i=1}^{N}\langle Z_{ij}\rangle\,, \quad \vartheta_j = \langle\Psi_j\rangle + \sum_{1}^{N}\sum_{s=j+1}^{M}\langle Z_{is}\rangle \tag{33}
$$

$$
a_j^* = a_j + 1\,, \quad b_j^* = b_j - \langle\ln(1-\lambda_j)\rangle \tag{34}
$$

$$
\rho_k = 1 + \sum_{i=1}^{N}\sum_{l=1}^{D}\langle W_{ikl}\rangle\,, \quad \varpi_k = \langle\varphi_k\rangle + \sum_{i=1}^{N}\sum_{s=k+1}^{K}\sum_{l=1}^{D}\langle W_{isl}\rangle \tag{35}
$$

$$
c_k^* = c_k + 1\,, \quad d_k^* = d_k - \langle\ln(1-\gamma_k)\rangle \tag{36}
$$

$$
\xi_1^* = \xi_1 + \sum_{i=1}^{N}\langle\phi_{il}\rangle\,, \qquad \xi_2^* = \xi_2 + \sum_{i=1}^{N}\langle 1-\phi_{il}\rangle \tag{37}
$$

$$
\begin{aligned}
u_{jl}^* = & \, u_{jl} + \sum_{i=1}^{N}\langle Z_{ij}\rangle\langle\phi_{il}\rangle\left[\psi\left(\bar{\alpha}_{jl}+\bar{\beta}_{jl}\right) - \psi\left(\bar{\alpha}_{jl}\right)\right. \\
& \left. + \bar{\beta}_{jl}\psi'\left(\bar{\alpha}_{jl}+\bar{\beta}_{jl}\right)\right)(\langle\ln\beta_{jl}\rangle - \ln\bar{\beta}_{jl})\Big]\bar{\alpha}_{jl}
\end{aligned} \tag{38}
$$

$$v_{jl}^{*} = v_{jl} - \sum_{i=1}^{N}\langle Z_{ij}\rangle\langle\phi_{il}\rangle \ln\frac{X_{il}}{1+X_{il}} \tag{39}$$

$$\begin{aligned} p_{jl}^{*} &= p_{jl} + \sum_{i=1}^{N}\langle Z_{ij}\rangle\langle\phi_{il}\rangle\left[\psi\left(\bar{\beta}_{jl}+\bar{\alpha}_{jl}\right) - \psi\left(\bar{\beta}_{jl}\right)\right. \\ &\quad \left. + \bar{\alpha}_{jl}\psi'\left(\bar{\beta}_{jl}+\bar{\alpha}_{jl}\right)\right)\left(\langle\ln\alpha_{jl}\rangle - \ln\bar{\alpha}_{jl}\right)\Big]\bar{\beta}_{jl} \end{aligned} \tag{40}$$

$$q_{jl}^{*} = q_{jl} - \sum_{i=1}^{N}\langle Z_{ij}\rangle\langle\phi_{il}\rangle \ln\frac{1}{1+X_{il}} \tag{41}$$

$$\begin{aligned} g_{kl}^{*} &= g_{kl} + \sum_{i=1}^{N}\langle 1-\phi_{il}\rangle\langle W_{ikl}\rangle\left[\psi\left(\bar{\sigma}_{kl}+\bar{\tau}_{kl}\right) - \psi\left(\bar{\sigma}_{kl}\right)\right. \\ &\quad \left. + \bar{\tau}_{kl}\psi'\left(\bar{\sigma}_{kl}+\bar{\tau}_{kl}\right)\right)\left(\langle\ln\tau_{kl}\rangle - \ln\bar{\tau}_{kl}\right)\Big]\bar{\sigma}_{kl} \end{aligned} \tag{42}$$

$$h_{kl}^{*} = h_{kl} - \sum_{i=1}^{N}\langle 1-\phi_{il}\rangle\langle W_{jkl}\rangle \ln\frac{X_{il}}{1+X_{il}} \tag{43}$$

$$\begin{aligned} s_{kl}^{*} &= s_{kl} + \sum_{i=1}^{N}\langle 1-\phi_{il}\rangle\langle W_{ikl}\rangle\left[\psi\left(\bar{\tau}_{kl}+\bar{\sigma}_{kl}\right) - \psi\left(\bar{\tau}_{kl}\right)\right. \\ &\quad \left. + \bar{\sigma}_{kl}\psi'\left(\bar{\tau}_{kl}+\bar{\sigma}_{kl}\right)\right)\left(\langle\ln\sigma_{kl}\rangle - \ln\bar{\sigma}_{kl}\right)\Big]\bar{\tau}_{kl} \end{aligned} \tag{44}$$

$$t_{kl}^{*} = t_{kl} - \sum_{i=1}^{N}\langle 1-\phi_{il}\rangle\langle W_{jkl}\rangle \ln\frac{1}{1+X_{il}} \tag{45}$$

The expected values are given by

$$\bar{\alpha}_{jl} = \frac{u_{jl}^{*}}{v_{jl}^{*}}, \quad \bar{\beta}_{jl} = \frac{p_{jl}^{*}}{q_{jl}^{*}}, \quad \bar{\sigma}_{kl} = \frac{g_{kl}^{*}}{h_{jl}^{*}}, \quad \bar{\tau}_{kl} = \frac{s_{kl}^{*}}{t_{kl}^{*}} \tag{46}$$

$$\langle\psi_j\rangle = \frac{a_j^{*}}{b_j^{*}}, \quad \langle\varphi_k\rangle = \frac{c_k^{*}}{d_k^{*}}, \quad \langle Z_{ij}\rangle = r_{ij}, \quad \langle W_{ikl}\rangle = m_{ikl} \tag{47}$$

$$\langle\phi_{il}\rangle = f_{il}, \quad \langle 1-\phi_{il}\rangle = 1-f_{il}, \quad \langle\ln\alpha\rangle = \psi(u^{*}) - \ln v^{*}, \quad \langle\ln\beta\rangle = \psi(p^{*}) - \ln q^{*} \tag{48}$$

$$\langle\ln\sigma\rangle = \psi(g^{*}) - \ln h^{*}, \quad \langle\ln\tau\rangle = \psi(s^{*}) - \ln t^{*} \tag{49}$$

$$\langle\ln\lambda_j\rangle = \psi(\theta) - \psi(\theta+\vartheta), \quad \langle\ln(1-\lambda_j)\rangle = \psi(\vartheta) - \psi(\theta+\vartheta) \tag{50}$$

$$\langle\ln\gamma_k\rangle = \psi(\rho) - \psi(\rho+\varpi), \quad \langle\ln(1-\gamma_k)\rangle = \psi(\varpi) - \psi(\rho+\varpi), \tag{51}$$

$$\langle\ln\epsilon_{l_1}\rangle = \psi(\xi_1^{*}) - \psi(\xi_1^{*}+\xi_2^{*}), \quad \langle\ln\epsilon_{l_2}\rangle = \psi(\xi_2^{*}) - \psi(\xi_1^{*}+\xi_2^{*}) \tag{52}$$

$$\langle(\ln\alpha - \ln\bar{\alpha})^2\rangle = [\psi(u^{*}) - \ln u^{*}]^2 + \psi'(u^{*}) \tag{53}$$

$$\langle(\ln\beta - \ln\bar{\beta})^2\rangle = [\psi(p^{*}) - \ln p^{*}]^2 + \psi'(p^{*}) \tag{54}$$

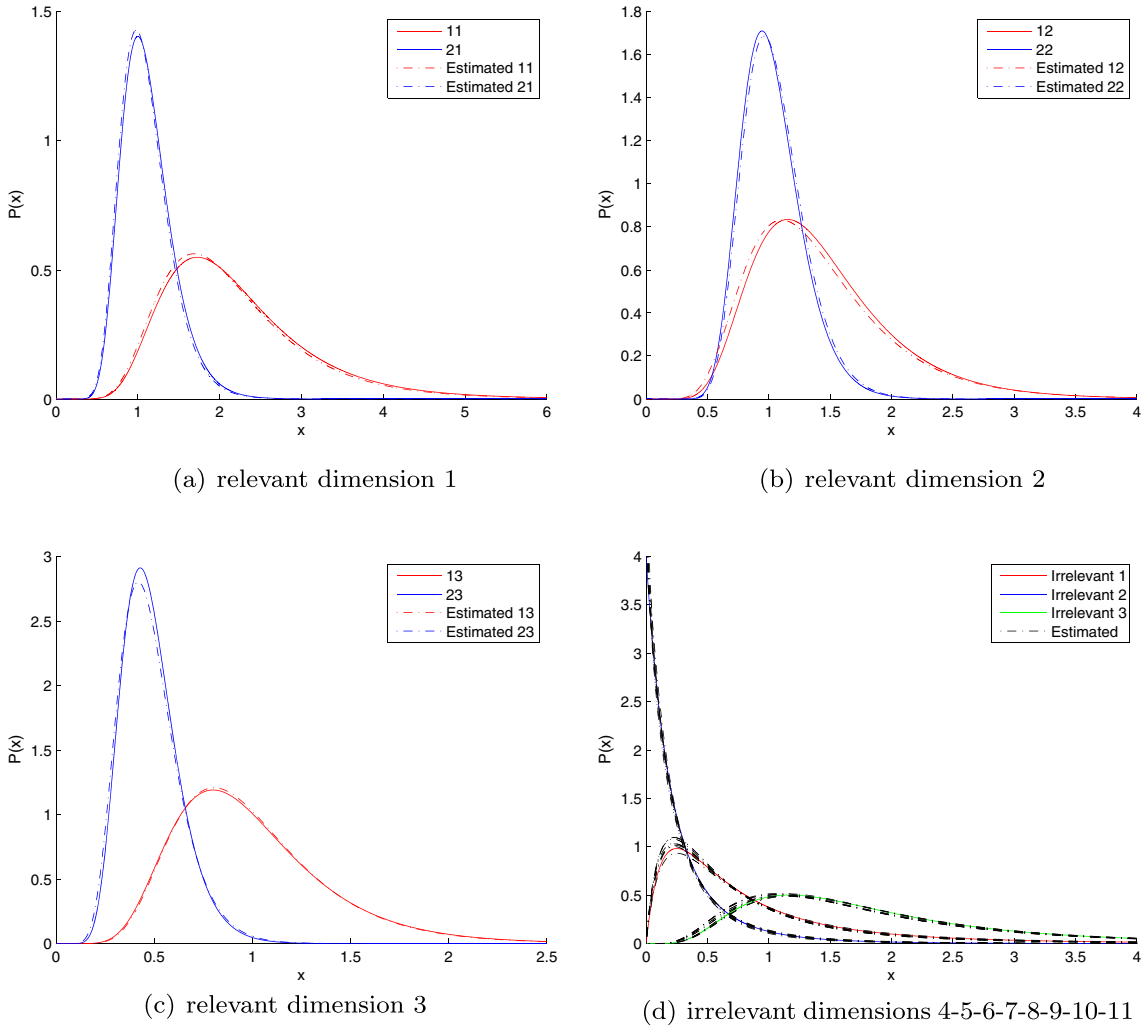
$$\langle(\ln\sigma - \ln\bar{\sigma})^2\rangle = [\psi(g^{*}) - \ln g^{*}]^2 + \psi'(g^{*}) \tag{55}$$

$$\langle(\ln\tau - \ln\bar{\tau})^2\rangle = [\psi(s^{*}) - \ln s^{*}]^2 + \psi'(s^{*}) \tag{56}$$

In order to initialize M and K, we adopt the same strategy adopted by [38, 52, 53] which consists of over-initializing the number of clusters M and K to be much larger than the true model, and then infer the structure of the mixture by discarding the components whose mixing probabilities are close to zero. The learning process is summarized in Algorithm 1.

Algorithm 1 Infinite GID learning using variational inference with feature selection

1: Choose truncations levels M and K
2: Initialize the values for hyperparameters $u_{ji}, v_{ji}, p_{ji}, q_{ji}, g_{kl}, h_{kl}, s_{kl}, t_{kl}, a_j, b_j, c_k, d_k, \xi_1$, and ξ_2.
3: Initialize the values of r_{ij} and m_{ikl} using the K-means algorithm.
4: Estimate the expected value using Eqs.46-56.
5: Update the variational solutions of each factor using Eqs.25.
6: If convergence criteria is reached go to 7. else go to 4.
7: Compute the expected value of λ_j as $\langle\lambda_j\rangle = \frac{\theta_j}{\theta_j+\vartheta_j}$ and substitute it into Eq.11 in order to compute the estimated mixing probabilities π_j
8: Compute the expected value of γ_k as $\langle\gamma_k\rangle = \frac{\rho_k}{\rho_k+\varpi_k}$ and substitute it into Eq.11 in order to compute the estimated mixing probabilities η_k
9: Calculate the expected values of features saliencies $\langle\epsilon_l\rangle = \frac{\xi_1^{*}}{\xi_1^{*}+\xi_2^{*}}$
10: Set the optimal number of components M and K by discarding the components whose mixing probabilities are close to 0.



(a) relevant dimension 1

(b) relevant dimension 2

(c) relevant dimension 3

(d) irrelevant dimensions 4-5-6-7-8-9-10-11

Fig. 2 Different generated inverted Beta distributions labeled in ji representing the 11 dimensions of the model in Table 1

4 Experimental results

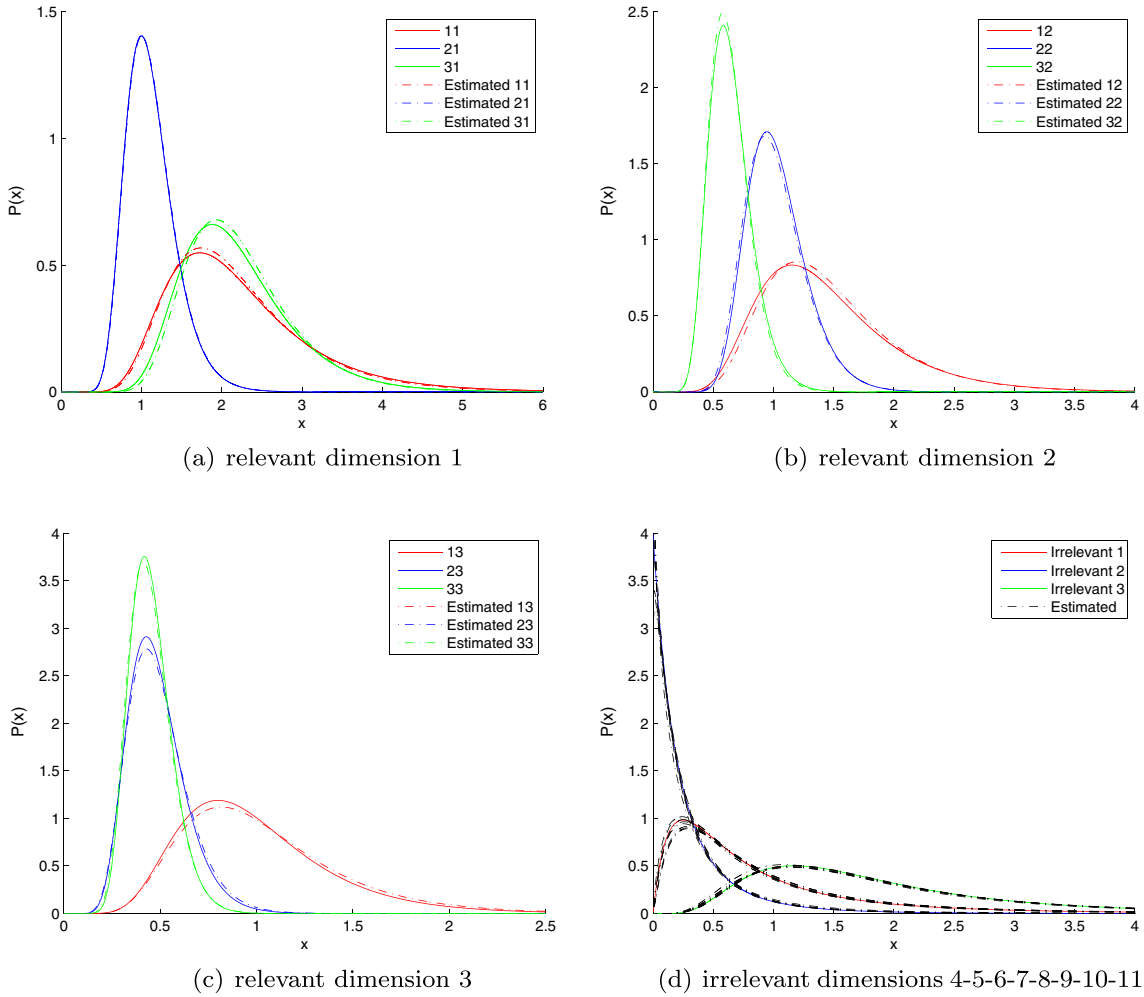
In this section, an evaluation of our proposed algorithm is performed using synthetic and real computer vision datasets. The values of hyper parameters are empirically initialized basing on several runs. u, p, h, t, g and s are set to 1, and v, q are set to 0.05. The hyper parameters a, b, c and d are set to 1, and ξ_1, ξ_2 are set to 0.01. The choice of these parameters have been used in all our experiments. The hyper parameters can be randomly set, but a good initialization is still crucial because the Kullback-Leibler divergence to the true posterior may contain many local minima [54]. Several methods have been proposed to estimate an initial value for the hyperparameters in [54, 55] and will not be discussed further.

Table 1 Real and estimated parameters of relevant components in the case of a 11-dimensional dataset generated from 2-components mixture model

j	l	n_j	π_j	α_{ji}	β_{jl}	$\hat{\pi}_j$	$\hat{\alpha_{jl}}$	$\hat{\beta_{jl}}$
1	1	600	0.5	20	10	0.5077	19.76	10.11
	2			16	12		14.53	11.24
	3			13	14		13.48	14.50
2	1	600	0.5	28	26	0.4923	27.55	26.09
	2			35	35		35.60	34.97
	3			16	34		14.29	30.71

4.1 Synthetic data

At a first stage we evaluate the performance of our algorithm using synthetic data. We propose to use 11-dimensional datasets, whose first three dimensions are relevant and the remaining eight dimensions are irrelevant. Depending on



(a) relevant dimension 1 (b) relevant dimension 2

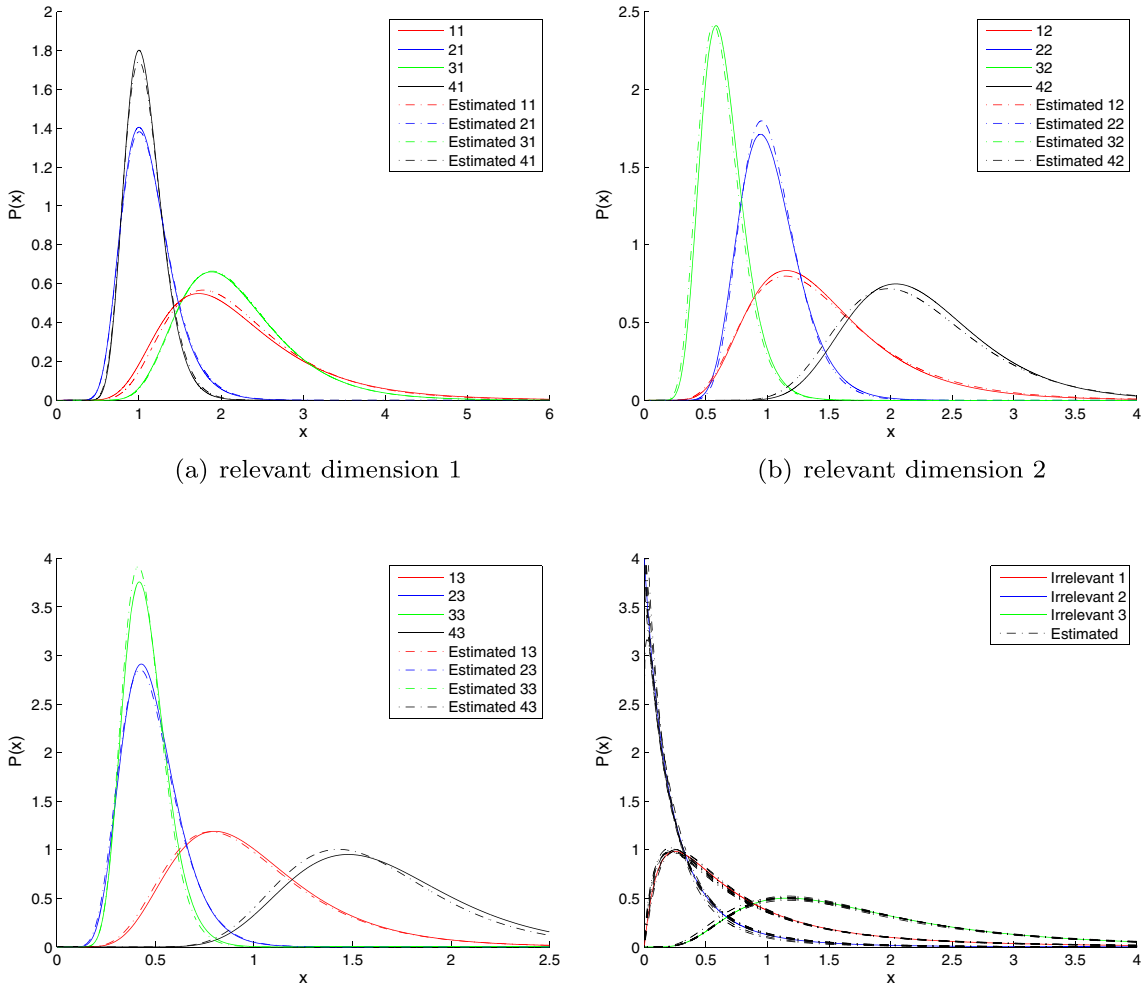
(c) relevant dimension 3 (d) irrelevant dimensions 4-5-6-7-8-9-10-11

Fig. 3 Different generated inverted Beta distributions labeled in ji representing the 11 dimensions of the model in Table 2

the dataset, the relevant features are generated using at least 2 distinguishable inverted beta distributions, while the irrelevant features are generated by a mixture whose inverted beta distributions are all overlapping ($iBeta(2,3)$, $iBeta(1,4)$, $iBeta(8,5)$). For all our experiments on synthetic data we use an initial value of $M = 15$ and $K = 10$. The first dataset consists of two components, composed of 600 samples each. Figure 2 represents the real and estimated histograms, where Fig. 2a, b and c represent the different inverted beta distributions for the first three relevant dimensions while Fig. 2d represents the distributions of the 8 irrelevant features. The continuous lines represent the real histograms while the dashed ones represent the estimated histograms. The obtained results are reported in Table 1 where we show the real and estimated parameters. For ease of presentation we do not illustrate the values of the estimated parameters of the irrelevant features, but as it can be shown in Fig. 2d, the estimated histograms reflect a good estimation. The algorithm was capable to find the correct number of classes $M = 2$ and the correct number of distributions forming the irrelevant features which is $K = 3$.

Table 2 Real and estimated parameters of relevant components in the case of a 11-dimensional dataset generated from 3-components mixture model

j	l	n_j	π_j	α_{ji}	β_{jl}	$\hat{\pi}_j$	$\hat{\alpha_{jl}}$	$\hat{\beta_{jl}}$
1	1	600	0.4	20	10	0.3988	21.61	10.81
	2			16	12		17.94	13.19
	3			13	14		12.34	12.88
2	1	600	0.4	28	26	0.4084	28.22	26.10
	2			35	35		32.68	33.11
	3			16	34		14.92	31.34
3	1	300	0.2	33	16	0.1928	36.89	17.56
	2			22	35		22.99	37.02
	3			24	54		22.64	51.48



(a) relevant dimension 1

(b) relevant dimension 2

(c) relevant dimension 3

(d) irrelevant dimensions 4-5-6-7-8-9-10-11

Fig. 4 Different generated inverted Beta distributions labeled in ji representing the 11 dimensions of the model in Table 3

The feature saliency was also correctly estimated, as $\epsilon_{l1} = 1$ for the first three features and zero for the rest. The maximum detected error is 10.69 %, which remains a good estimate with the presence of irrelevant features and the fact that we have considered slightly overlapping relevant features.

The second dataset is constructed by adding a third component to the first dataset, as shown in Fig. 3, where we plot the estimated and real histograms of the different dimensions, as for the first dataset in Fig. 2. The variational inference algorithm was capable to find the correct number of classes $M = 3$ and $K = 3$. We report on the estimated parameters in Table 2, where the maximum detected relative error is equal to 12.13 %. Still the estimated histograms are almost indistinguishable from the real ones, which means that the estimated parameters lead to the same performance of clustering as the real parameters.

Table 3 Real and estimated parameters of relevant components in the case of a 11-dimensional dataset generated from 4-components mixture model

j	l	n_j	π_j	α_{ji}	β_{jl}	$\hat{\pi}_j$	$\hat{\alpha_{jl}}$	$\hat{\beta_{jl}}$
1	1	600	0.30	20	10	0.3053	22.53	11.09
	2			16	12		14.78	11.00
	3			13	14		12.29	13.45
2	1	600	0.30	28	26	0.2991	27.45	25.34
	2			35	35		38.99	38.92
	3			16	34		15.02	32.16
3	1	300	0.15	33	16	0.1447	33.66	16.28
	2			22	35		20.44	33.51
	3			24	54		25.20	57.62
4	1	500	0.25	44	42	0.2509	41.42	39.44
	2			50	23		43.15	20.26
	3			35	22		35.52	23.16

Fig. 5 Visual Scenes Dataset : Forest, InsideCity, Highway, TallBuilding

The third dataset that we used was constructed by adding a fourth component to the third dataset as illustrated in Fig. 4. The algorithm was capable of finding the correct number of relevant components $M = 4$, and the number of components composing the mixture of irrelevant features $K = 3$. We report on the estimated parameters of the relevant features in Table 3 where the maximum detected relative error is equal to 13.70 %. It is noteworthy that the synthetic dataset that we are considering are difficult to estimate even with the absence of irrelevant features as the relevant features are still overlapping. The obtained results show that our algorithm performs well for synthetic data, as it was capable to optimally select M and K, and generate estimated histograms that are almost indistinguishable from the real ones, leading to the same clustering results.

4.2 Visual scene categorization

In this section we investigate the performance of our algorithm using real-life data. We consider the publicly available visual scenes dataset that was proposed in [56]. We mainly used four categories which are Highway (260 images), InsideCity (308 images), TallBuilding (356 images), and Forest (328 images). Figure 5 illustrates samples of the considered dataset. The images within a given category are diverse and the visual scenes have different objects, colors and shapes. The features of each image have been extracted using the local Histogram of Oriented Gradient (HOG) descriptor proposed in [57]. The HOG algorithm is efficient in terms of detecting local characteristics of images and object recognition. During our experiments each image was represented by a 81-dimensional feature vector.

The experiments consists of establishing a clustering of the data basing on the hierarchal construction strategy that has been proposed in [17], where a parent cluster is composed of sub children clusters. As proposed by [17], we label each cluster according to the images that have been grouped into it. Indeed, we attribute each cluster k to the super class j whose elements are the most present in cluster k such that

$$label_{cluster_k} = \arg\max_j \frac{elements\ of\ superclass\ j\ in\ cluster_k}{elements\ in\ cluster_k} \quad (57)$$

We consider that we have four super classes that are Highway, InsideCity, TallBuiling and Forest. Indeed, it has

Table 4 Visual Scenes Var-InfGID-FS confusion matrix

	Forest	Highway	InsideCity	TallBuilding
Forest	**324**	0	0	4
Highway	11	**238**	9	2
InsideCity	22	9	**192**	85
TallBuilding	28	13	25	**290**

Table 5 Visual Scenes Var-InfGID confusion matrix

	Forest	Highway	InsideCity	TallBuilding
Forest	**322**	0	2	4
Highway	1	**194**	64	1
InsideCity	0	9	**184**	115
TallBuilding	1	5	46	**304**

Table 6 Visual Scenes MML-EM-GID confusion matrix

	Forest	Highway	InsideCity	TallBuilding
Forest	**319**	0	2	7
Highway	0	**188**	70	2
InsideCity	0	5	**246**	57
TallBuilding	1	3	116	**236**

been shown in [17] that it is more appropriate to represent a given object class by a mixture than one single distribution. This is convenient with variational inference, since it is able to set the optimal number of components to model a given dataset without the need of forcing the number of components to be equal to the number of object classes. Thus, the optimal number of components can be largely higher than the number of object classes. The work in [34] particularly shows that the appropriate number of components for the mixture can be determined in a single training run without recourse to cross-validation using variational inference. In this experiment we consider to compare mixture models that mainly use the inverted Dirichlet distributions and the generalized inverted Dirichlet distributions that involve inverted Beta distributions for each dimension. It as been shown in previous works that these mixture models outperform the GMM in terms of clustering accuracy [16, 18]. We propose to consider the variational infinite GID mixture with feature selection (Var-InfGID-FS), the variational infinite GID mixture without considering feature selection (Var-InfGID), the GID mixture using EM and MML (MML-EM-GID), and the ID mixture using EM and MML (MML-EM-ID). For the Var-InfGID-FS we set the initial values of $M = 25$ and $K = 30$, following the strategy proposed by [38, 52, 53], which consists of setting a larger number of clusters than four in our case. For the Var-InfGID we set the initial values of $M = 25$, and as for the MML-EM-GID and MML-EM-GID, we consider 25 models going from 1 component to 25 components, and then we use the MML as defined in [16, 18] to select the optimal model. Tables 4, 5, 6 and 7 show the confusion matrices of the clustering results using Var-InfGID-FS, Var-InfGID, MML-EM-GID and MML-EM-ID, respectively. Using those confusion matrices we calculate the clustering accuracies reported in Table 8, with the relevant number of clusters that constitute the model. The obtained results show that the Var-InfGID-FS outperforms the other algorithms with an accuracy equal to 83.39 %, followed by the Var-InfGID with accuracy equal to 80.19 % which shows the merit of variational inference and feature selection.

4.3 Digits categorization

For the second application on a real computer vision dataset, we consider the public available MNIST digits database which is a large database of handwritten digits [58]. Each MNIST image is a digitized picture of a single handwritten digit character. Each image is 28 x 28 pixels in size. Each pixel value is between 0, which represents white, and 255, which represents black. Intermediate pixel values represent shades of gray. Figure 6 shows samples from 1 to 9 of the handwritten digits. Digits recognition may be easy for a human being in most of cases, but it still is a challenging application for machine learning approaches. We also consider the HOG features that we extract from 60000 digits images that are distributed as follows; Zero (5923 images), One (6742 images), Two (5958 images), Three (6131 images), Four (5842 images), Five (5421 images), Six (5918 images), Seven (6265 images), Eight (5851 images), Nine (5949 images). Bear in mind that our experiment consists of clustering without any training phase as data is injected to the algorithm without any prior knowledge about the observations labels. As illustrated in Fig. 6, the optical digits for a given number take different shapes and orientations, therefore we expect them to form a considerable number of clusters for one single number. We set an initial value of $M = 500$ for the Var-InfGID-FS and Var-InfGID. We set $K = 100$ for the Var-InfGID-FS. The use of MML is not practical in this case, especially when the number of classes increases significantly and we have to go through 500 models that are computationally costly in order to select the best model, in contrast with the variational inference that needs a single run in order to select the optimal value of M. Therefore this experiment is restricted on the Var-InfGID-FS and Var-InfGID, in order to show the impact of feature selection on the clustering accuracy. Tables 9 and 10

Table 7 Visual Scenes MML-EM-ID confusion matrix

	Forest	Highway	InsideCity	TallBuilding
Forest	**320**	0	2	6
Highway	1	**185**	74	0
InsideCity	0	5	**281**	22
TallBuilding	1	4	152	**199**

Table 8 Obtained Accuracies and corresponding of clusters for each model : Visual Scenes Dataset

	GID VAR FS	GID VAR	GID EM MML	ID EM MML
Accuracy	83.39 %	80.19 %	78.99 %	78.67 %
Number of Clusters	7	5	6	8

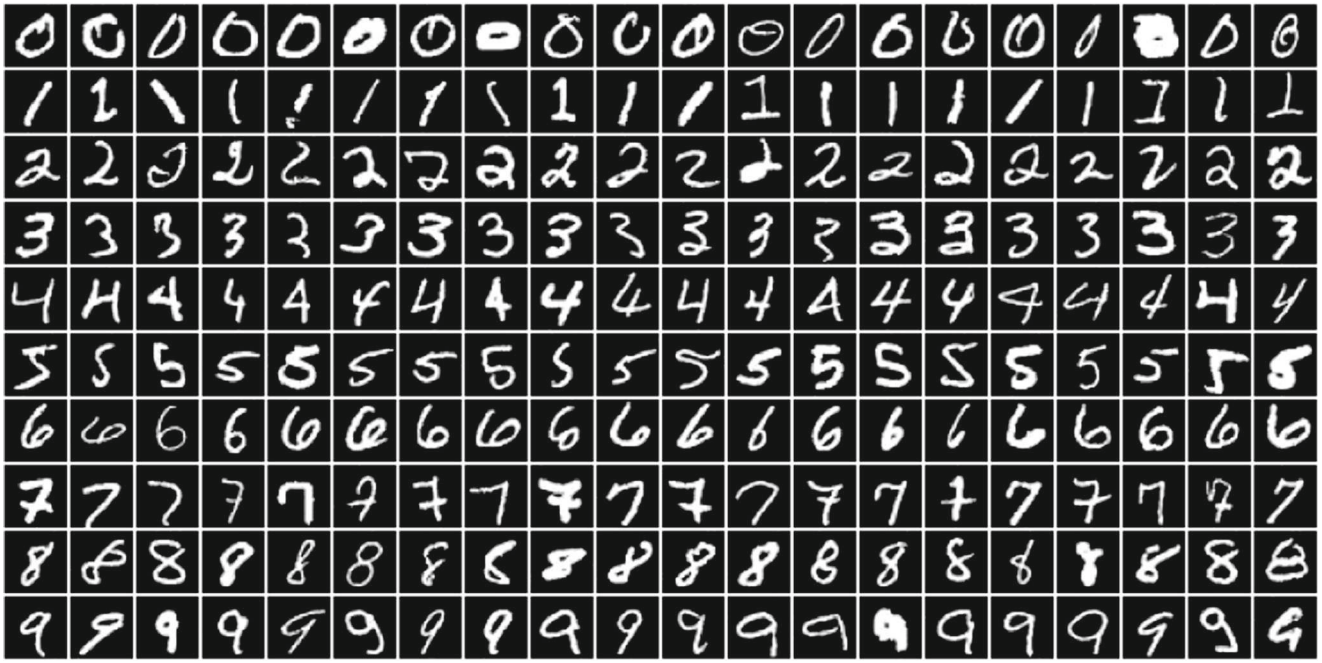


Fig. 6 Digits dataset : sample images

respectively show the confusion matrices obtained using the Var-InfGID-FS and Var-InfGID. We report on the accuracy of clustering in Table 11.

We notice that the accuracy is improved by 0.96 % when using feature selection, which concern 576 images out of 60000. Naturally the impact of this accuracy improvement increases when the number of clustered data is considerable.

4.4 Discussion

The model we have built shows the capability of the variational inference over the EM learning using MML. Using the Var-InfGID-FS improved the clustering accuracy by 4.72 % (using MML-EM-ID), 4.40 % (using MML-EM-GID), and 3.20 % (using Var-InfGID), when applied on the visual scenes dataset. Also, using the Var-InfGID-FS improved the accuracy of clustering by 0.96 % (using Var-InfGID) when applied on the 60000 digits dataset, which shows the merit of feature selection. Figure 7 show the features saliencies of the two datasets when using the Var-InfGID-FS. We notice that the algorithm was capable to assign different weights to features as some features are insignificant in the clustering process, and some others have a high discriminative effect. When we compare the number of clusters obtained when using the Var-InfGID-FS with the one obtained when using Var-InfGID, we notice that

Table 9 Digits Var-InfGID-FS confusion matrix

	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Zero	**5716**	39	22	4	2	18	23	11	67	21
One	0	**6522**	64	1	6	0	4	16	28	101
Two	12	14	**5502**	198	15	13	4	83	96	21
Three	16	12	436	**5036**	6	229	2	125	202	67
Four	1	45	23	2	**5320**	5	88	26	46	286
Five	31	26	20	209	7	**4712**	65	32	238	81
Six	31	57	9	1	23	40	**5655**	0	100	2
Seven	6	23	96	48	2	4	0	**5753**	43	290
Eight	38	59	61	70	18	192	97	43	**5128**	145
Nine	63	29	24	49	62	27	11	229	103	**5352**

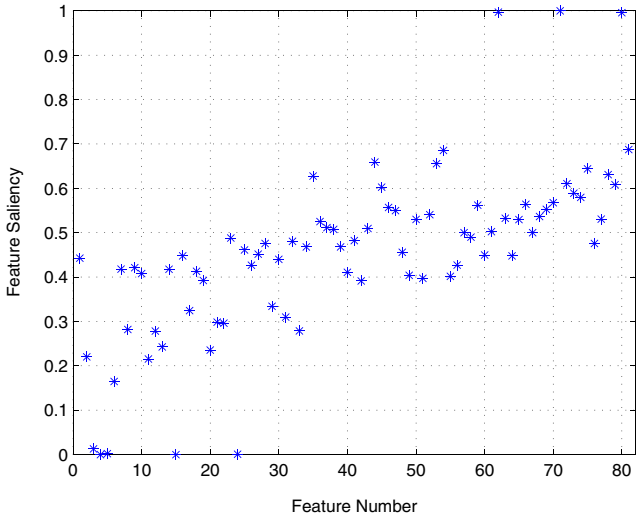
Table 10 Digits Var-InfGID confusion matrix

	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Zero	**5601**	47	18	128	0	26	36	13	46	8
One	0	**6595**	92	10	5	1	3	32	0	4
Two	6	15	**5413**	374	2	8	7	67	49	17
Three	9	3	373	**5431**	3	130	4	61	72	45
Four	0	51	25	21	**5396**	11	121	32	25	160
Five	13	4	24	545	2	**4553**	76	12	159	33
Six	34	59	12	25	16	25	**5678**	1	66	2
Seven	0	9	101	151	20	22	1	**5868**	14	79
Eight	49	59	48	497	29	191	147	44	**4640**	147
Nine	53	43	20	318	127	16	11	355	64	**4942**

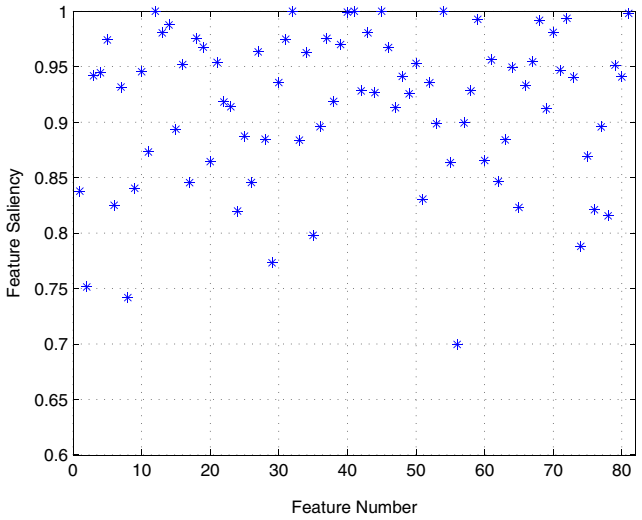
Table 11 Obtained accuracies and corresponding of clusters for each model: Digits Dataset

	GID VAR FS	GID VAR
Accuracy	91.16 %	90.20 %
Number of Clusters	240	179

usually the number of clusters resulted by the Var-InfGID-FS is higher. This is due to the fact that feature selection leads to a denoising process that discard the features that do not have a discriminative effect and focus on the features that contribute the most to construct clusters with a higher cohesion which may lead to an increase of the number of clusters.



(a) Visual Scenes dataset



(b) Digits dataset

Fig. 7 Features saliencies of different datasets

5 Conclusion

In this work we have proposed an approach to model and cluster data using an infinite GID mixture with feature selection. The model integrates a set of inverted Beta distributions that represent two majors aspects; relevant features and irrelevant features. The variational Bayesian inference has been used to estimate the parameters of our model, and the obtained results when applied on real data show its merit over the conventional EM algorithm. We have also shown that the use of feature selection leads to better results in terms of clustering accuracy. The proposed model can be used for any positive data and has promising applications in different areas that have huge amount of data to be clustered and analyzed statistically.

Acknowledgments The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC); and Concordia University via a Research Chair in Management, Analysis, and Modeling of Big Multimodal Data and Applications. The authors would like to thank the anonymous referees and the associate editor for their helpful comments.

Appendix A: Conditional independence in the transformed space

We know that the posterior probability is $p(j|\mathbf{Y}_i) \propto \pi_j p(\mathbf{Y}_i|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$, so every vector $\mathbf{Y}_i$ is assigned to its cluster j such as $j = \arg\max_j p(j|\mathbf{Y}_i) = \arg\max_j \pi_j p(\mathbf{Y}_i|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$. We have:

$$p(\mathbf{Y}_i|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) = \prod_{d=1}^{D} \frac{\Gamma(\alpha_{jd}+\beta_{jd})}{\Gamma(\alpha_{jd})\Gamma(\beta_{jd})} \frac{Y_{id}^{\alpha_{jd}-1}}{(1+\sum_{l=1}^{d} Y_{il})^{\gamma_{jd}}} \tag{58}$$

For GID, it is possible to compute the posterior probability by examining the form of the product in (58) and considering every feature separately, so if we want to consider the feature D, (58) becomes for a specific vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iD})$:

$$\frac{1}{B(\alpha_{jD}, \beta_{jD})} Y_{iD}^{\alpha_{jD}-1} \left(1+\sum_{l=1}^{D} Y_{il}\right)^{-\beta_{jD}-\alpha_{jD}+\beta_{j(D+1)}} \times \prod_{l=1}^{D-1} \frac{1}{B(\alpha_{jl}, \beta_{jl})} Y_{iD}^{\alpha_{jl}-1} \left(1+\sum_{k=1}^{l} Y_{ik}\right)^{-\beta_{jl}-\alpha_{jl}+\beta_{j(l+1)}} \tag{59}$$

where $B(\alpha_{jl},\beta_{jl})$ is the beta function such that $B(\alpha_{jl},\beta_{jl})=\frac{\Gamma(\alpha_{jl})\Gamma(\alpha_{jl})}{\Gamma(\alpha_{jl}+\beta_{jl})}$. As $\beta_{j(D+1)}=0$, (59) becomes:

$$\frac{1}{B(\alpha_{jD},\beta_{jD})}Y_{iD}^{\alpha_{jD}-1}\left(1+\sum_{l=1}^{D}Y_{il}\right)^{-\beta_{jD}-\alpha_{jD}}\prod_{l=1}^{D-1}\times\frac{1}{B(\alpha_{jl},\beta_{jl})}Y_{iD}^{\alpha_{jl}-1}\left(1+\sum_{k=1}^{l}Y_{ik}\right)^{-\beta_{jl}-\alpha_{jl}+\beta_{j(l+1)}} \tag{60}$$

by multiplying (60) by the constant $\left(1+\sum_{l=1}^{D-1}Y_{il}\right)^{\beta_{jD}+\alpha_{jD}-\alpha_{jD}+1}=\left(1+\sum_{l=1}^{D-1}Y_{il}\right)^{\beta_{jD}+1}$, (60) becomes proportional to:

$$\frac{1}{B(\alpha_{jD},\beta_{jD})}\left(\frac{Y_{iD}}{1+\sum_{l=1}^{D-1}Y_{il}}\right)^{\alpha_{jD}-1}\left(1+\frac{Y_{iD}}{1+\sum_{l=1}^{D-1}Y_{il}}\right)^{-\beta_{jD}-\alpha_{jD}}\times\prod_{l=1}^{D-1}\frac{1}{B(\alpha_{jl},\beta_{jl})}Y_{iD}^{\alpha_{jl}-1}\left(1+\sum_{k=1}^{l}Y_{ik}\right)^{-\beta_{jl}-\alpha_{jl}+\beta_{j(l+1)}} \tag{61}$$

We know that:

$$\frac{1}{B(\alpha_{jD},\beta_{jD})}\left(\frac{Y_{iD}}{1+\sum_{l=1}^{D-1}Y_{il}}\right)^{\alpha_{jD}-1}\left(1+\frac{Y_{iD}}{1+\sum_{l=1}^{D-1}Y_{il}}\right)^{-\beta_{jD}-\alpha_{jD}}=p_{iBeta}\left(\frac{Y_{iD}}{1+\sum_{l=1}^{D-1}Y_{il}}|\alpha_{jD},\beta_{jD}\right) \tag{62}$$

so (60) becomes:

$$p_{iBeta}\left(\frac{Y_{iD}}{1+\sum_{l=1}^{D-1}Y_{il}}|\alpha_{jD},\beta_{jD}\right)\prod_{l=1}^{D-1}\frac{1}{B(\alpha_{jl},\beta_{jl})}Y_{iD}^{\alpha_{jl}-1}\times\left(1+\sum_{k=1}^{l}Y_{ik}\right)^{-\beta_{jl}-\alpha_{jl}+\beta_{j(l+1)}} \tag{63}$$

For every remaining feature l in the product from 1 to $D-1$ we mutiply (63) by the constant $\left(1+\sum_{k=1}^{l-1}Y_{ik}\right)^{\beta_{jl}+\alpha_{jl}-\alpha_{jl}+1}\left(1+\sum_{k=1}^{l}Y_{ik}\right)^{-\beta_{j(l+1)}}=\left(1+\sum_{k=1}^{l-1}Y_{ik}\right)^{\beta_{jl}+1}\left(1+\sum_{k=1}^{l}Y_{ik}\right)^{-\beta_{j(l+1)}}$ so (63) will be proportional to:

$$\prod_{l=1}^{D}p_{iBeta}\left(\frac{Y_{il}}{1+\sum_{k=1}^{l-1}Y_{ik}}|\alpha_{jl},\beta_{jl}\right) \tag{64}$$

the first term of the product in (64) is : $p_{iBeta}(Y_{i1}|\alpha_{jl},\beta_{jl})$ so we finally have:

$$p(j|\mathbf{Y}_i)\propto\pi_j p(\mathbf{Y}_i|\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)\propto\pi_j p_{iBeta}(Y_{i1}|\alpha_{jl},\beta_{jl})\times\prod_{l=2}^{D}p_{iBeta}\left(\frac{Y_{il}}{1+\sum_{k=1}^{l-1}Y_{ik}}|\alpha_{jl},\beta_{jl}\right) \tag{65}$$

Appendix B: Proof of equations

Equation (21) shows that the terms that are independent of $Q_s(\Theta_s)$ are absorbed into an additive constant. In order to make use of (21) we need to calculate the logarithm of (18) with the truncation of number of components of the GID mixture M, and the number of components of the irrelevant features to K. We also know that $\left\{\sum_{j=1}^{M}\langle Z_{ij}\rangle\right\}=1$, so this term will be discarded when it is factorized in the variational factors.

Variational solution to $Q(\boldsymbol{\phi})$

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

$$\begin{aligned}\ln Q(\phi_{il}) \;=\; & \phi_{il}\left\{\langle\ln\epsilon_{l1}\rangle+\sum_{j=1}^{M}\langle Z_{ij}\rangle\left[\mathcal{R}_{jl}+(\bar{\alpha}_{jl}-1)\ln X_{il}\right.\right.\\ & \left.\left.-(\bar{\alpha}_{jl}+\bar{\beta}_{jl})\ln(1+X_{il})\right]\right\}\\ & +(1-\phi_{il})\left\{\langle\ln\epsilon_{l2}\rangle+\left\{\sum_{k=1}^{K}\langle W_{ikl}\rangle\left[\mathcal{F}_{kl}+(\bar{\sigma}_{kl}-1)\ln X_{il}\right.\right.\right.\\ & \left.\left.\left.-(\bar{\sigma}_{kl}+\bar{\tau}_{kl})\ln(1+X_{il})\right]\right\}\right\}+const\end{aligned} \tag{66}$$

where

$$\bar{\alpha}=\langle\alpha\rangle,\quad\bar{\beta}=\langle\beta\rangle,\quad\bar{\tau}=\langle\tau\rangle \tag{67}$$

and

$$\mathcal{R}_{jl}=\langle\ln\frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})}\rangle,\quad\mathcal{F}_{kl}=\langle\ln\frac{\Gamma(\sigma_{kl}+\tau_{kl})}{\Gamma(\sigma_{kl})\Gamma(\tau_{kl})}\rangle \tag{68}$$

The expectations in (68) are analytically intractable, thus, we apply the second-order Taylor series expansion in order to obtain a closed-form expression such as in [37]. The approximation of $\mathcal{R}_{jl}$ and $\mathcal{F}_{kl}$ are given by $\tilde{\mathcal{R}}_{jl}$ (31) and $\tilde{\mathcal{F}}_{kl}$

(32), respectively. We substitute the lower bound of (31) and (32) in (66) we obtain

$$\begin{aligned}\ln Q(\phi_{il}) \;=\; & \phi_{il}\left\{\langle \ln \epsilon_{l1}\rangle + \sum_{j=1}^{M}\langle Z_{ij}\rangle\left[\tilde{\mathcal{R}}_{jl} + (\bar{\alpha}_{jl}-1)\ln X_{il}\right.\right.\\ & \left.\left. - \;(\bar{\alpha}_{jl}+\bar{\beta}_{jl})\ln(1+X_{il})\right]\right\}\\ & +(1-\phi_{il})\left\{\langle \ln \epsilon_{l2}\rangle + \left\{\sum_{k=1}^{K}\langle W_{ikl}\rangle\left[\tilde{\mathcal{F}}_{kl}+(\bar{\sigma}_{kl}-1)\ln X_{il}\right.\right.\right.\\ & \left.\left.\left. - \;(\bar{\sigma}_{kl}+\bar{\tau}_{kl})\ln(1+X_{il})\right]\right\}\right\} + const \end{aligned} \tag{69}$$

From (69) we can deduce the variational solution of $Q(\boldsymbol{\phi})$ as a Bernoulli distribution such that

$$Q(\boldsymbol{\phi}) = \prod_{i=1}^{N}\prod_{l=1}^{D} f_{il}^{\phi_{il}}(1-f_{il})^{(1-\phi_{il})} \tag{70}$$

where f_{il} is defined in (26), and from the Bernoulli distribution it is straightforward to have

$$\langle \phi_{ij}\rangle = f_{ij}, \quad \langle 1-\phi_{ij}\rangle = 1-f_{ij} \tag{71}$$

Variational solution to Q(Z)

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

$$\begin{aligned}\ln Q(Z_{ij}) \;=\; & Z_{ij}\left\{\sum_{l=1}^{D}\left(\langle\phi_{il}\rangle\left[\tilde{\mathcal{R}}_{jl}+(\bar{\alpha}_{jl}-1)\ln X_{il} - (\bar{\alpha}_{jl}+\bar{\beta}_{jl})\ln(1+X_{il})\right]\right.\right.\\ & \left.\left. +\;\langle 1-\phi_{il}\rangle\sum_{k=1}^{K}\langle w_{ikl}\rangle\left[\tilde{\mathcal{F}}_{kl}+(\bar{\sigma}_{kl}-1)\ln X_{il} - (\bar{\sigma}_{kl}+\bar{\tau}_{kl})\ln(1+X_{il})\right]\right)\right.\\ & \left. +\;\langle \ln\lambda_j\rangle + \sum_{s=1}^{j-1}\langle\ln(1-\lambda_s)\rangle\right\} + const \end{aligned} \tag{72}$$

By analyzing the form of (72) we can write $\ln Q(Z)$ as

$$\ln Q(Z) = \sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}\ln\tilde{r}_{ij} + const \tag{73}$$

where $\tilde{r}_{ij}$ is defined in (27). Thus we have

$$Q(Z) \propto \prod_{i=1}^{N}\prod_{j=1}^{M}\tilde{r}_{ij}^{Z_{ij}} \tag{74}$$

We know that Z_{ij} are binary and we have $\sum_{j=1}^{M} Z_{ij} = 1$, so we can normalize (74) such that

$$Q(Z) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{75}$$

where r_{ij} is defined in (26). We can obtain $\langle Z_{ij}\rangle$ from the multinomial distribution of $Q(Z)$ such that

$$\langle Z_{ij}\rangle = r_{ij} \tag{76}$$

Variational solution to $Q(\boldsymbol{\lambda})$

The logarithm for of the variational factor $Q(\boldsymbol{\lambda})$ is given by

$$\begin{aligned}\ln Q(\lambda_j) \;=\; & \ln\lambda_j\sum_{i=1}^{N}\langle Z_{ij}\rangle + \ln(1-\lambda_j)\\ & \times\left(\sum_{1}^{N}\sum_{s=j+1}^{M}\langle Z_{is}\rangle + \langle\Psi_j\rangle - 1\right) + const \end{aligned} \tag{77}$$

Equation (77) has the logarithm form of the Beta distribution, by taking the exponential we obtain

$$Q(\boldsymbol{\lambda}) = \prod_{j=1}^{M} Beta(\lambda_j|\theta_j,\vartheta_j) \tag{78}$$

where θ_j and ϑ_j are defined in (33). As $\boldsymbol{\gamma}$ has the Beta prior distribution, $Q(\boldsymbol{\gamma})$ can be derived in a similar way as for $Q(\boldsymbol{\lambda})$. Following the same steps we define ρ_k and ϖ_k in (35)

Variational solution to $Q(\boldsymbol{\psi})$

The logarithm form of $Q(\boldsymbol{\psi})$ is given by

$$\ln Q(\psi_j) = \ln\psi_j a_j + \psi_j(\langle\ln(1-\lambda_j)\rangle - b_j) + const \tag{79}$$

by taking the exponential in "Variational solution to $Q(\boldsymbol{\psi})$" we obtain

$$Q(\boldsymbol{\psi}) = \prod_{j=1}^{M}\mathcal{G}(\psi_j|a_j^*,b_j^*) \tag{80}$$

where a_j^* and b_j^* are defined in (34). As $\boldsymbol{\varphi}$ has the Gamma prior distribution, $Q(\boldsymbol{\varphi})$ can be derived in a similar way as for $Q(\boldsymbol{\psi})$. Following the same steps we define c_k^* and d_k^* in (36).

Variational solution to $Q(W)$

The logarithm of the variational factor $Q(W_{ikl})$ is given by

$$\begin{aligned}\ln Q(W_{ikl}) \;=\; & W_{ikl}\left\{\langle 1-\phi_{il}\rangle\left(\tilde{\mathcal{F}}_{kl}+(\bar{\sigma}_{kl}-1)\ln X_{il}\right.\right.\\ & \left. - \;(\bar{\sigma}_{kl}+\bar{\tau}_{kl})\ln(1+X_{il})\right)\\ & \left. +\;\langle\ln\gamma_k\rangle + \sum_{s=1}^{k-1}\langle\ln(1-\gamma_s)\rangle\right\} + const \end{aligned} \tag{81}$$

By taking the exponential in (81) we obtain

$$Q(W) = \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{l=1}^{D} m_{ikl}^{W_{ikl}} \tag{82}$$

where m_{ikl} is given by (26).

Variational solution to $Q(\boldsymbol{\epsilon})$

The logarithm of the variational factor $Q(\boldsymbol{\epsilon}_l)$ is given by

$$\ln Q(\boldsymbol{\epsilon}_l) = \ln \epsilon_{l_1} \left(\sum_{i=1}^{N} \langle \phi_{il} \rangle + \xi_1 - 1 \right) + \ln \epsilon_{l_2} \left(\sum_{i=1}^{N} \langle 1 - \phi_{il} \rangle + \xi_2 - 1 \right) + const \quad (83)$$

Equation (83) has a logarithmic form similar to the logarithm form of a Dirichlet distribution. The variational solution to $Q(\boldsymbol{\epsilon}_l)$ can be obtained by

$$Q(\boldsymbol{\epsilon}_l) = \prod_{l=1}^{D} Dir(\boldsymbol{\epsilon}_l | \boldsymbol{\xi}^*) \quad (84)$$

where $\boldsymbol{\xi}^* = (\xi_1^*, \xi_2^*)$ in (37).

Variational solution to $Q(\boldsymbol{\alpha})$, $Q(\boldsymbol{\beta})$, $Q(\boldsymbol{\sigma})$, and $Q(\boldsymbol{\tau})$

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

$$\begin{aligned} \ln Q(\alpha_{jl}) &= \langle \ln p(\mathcal{X}, \Theta) \rangle_{\Theta \neq \alpha_{jl}} \\ &= \sum_{i=1}^{N} \langle Z_{ij} \rangle \langle \phi_{il} \rangle \left[\mathcal{D}(\alpha_{jl}) + \alpha_{jl} \ln \frac{X_{il}}{1 + X_{il}} \right] \\ &\quad + (u_{jl} - 1) \ln \alpha_{jl} - v_{jl} \alpha_{jl} + const \end{aligned} \quad (85)$$

where

$$\mathcal{D} = \langle \ln \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl}) \Gamma(\beta_{jl})} \rangle_{\beta_{jl}} \quad (86)$$

using a non-linear approximation as proposed in [37] we have

$$\mathcal{D}(\alpha) \geq \ln \alpha \left\{ \psi(\bar{\alpha} + \bar{\beta}) - \psi(\bar{\alpha}) + \bar{\beta} \psi'(\bar{\alpha} + \bar{\beta}) \right) \left(\langle \ln \beta \rangle - \ln \bar{\beta} \right) \Big\} \bar{\alpha} \quad (87)$$

We substitute the lower bound in (87) into the (85) we have

$$\begin{aligned} \ln Q(\alpha_{jl}) &= \ln \alpha_{jl} \left\{ \sum_{i=1}^{N} \langle Z_{ij} \rangle \langle \phi_{il} \rangle \left[\psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) \right. \right. \\ &\quad - \left. \psi(\bar{\alpha}_{jl}) + \bar{\beta}_{jl} \psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) \right) \left(\langle \ln \beta_{jl} \rangle - \ln \bar{\beta}_{jl} \right) \Big] \bar{\alpha}_{jl} \\ &\quad \left. + u_{jl} - 1 \right\} \\ &\quad + \alpha_{jl} \left\{ \sum_{i=1}^{N} \langle Z_{ij} \rangle \langle \phi_{il} \rangle \ln \frac{X_{il}}{1 + X_{il}} - v_{jl} \right\} + const \end{aligned} \quad (88)$$

Equation (88) has the form of a Gamma distribution. By taking it to the exponential we obtain

$$Q(\boldsymbol{\alpha}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\alpha_{jl} | u_{jl}^*, v_{jl}^*) \quad (89)$$

The hyperparameters u_{jl}^* and v_{jl}^* can be estimated by (38) and (39), respectively. Since $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ have the Gamma prior, we obtain the variational solutions to $Q(\boldsymbol{\beta})$, $Q(\boldsymbol{\sigma})$, and $Q(\boldsymbol{\tau})$ in the same way as for $Q(\boldsymbol{\alpha})$.

References

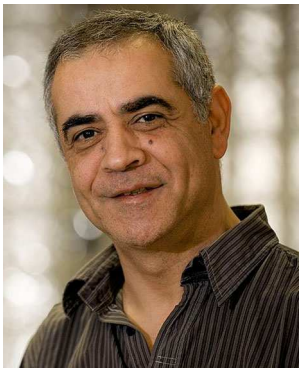
1. Jain AK, Murty M, Flynn P (1999) Data clustering: a Review. ACM Comput Surv 31(3):264–323
2. Rui X, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678
3. Bargary N, Hinde J, Garcia AF (2014) Finite mixture model clustering of snp data. In: MacKenziet G, Peng D (eds) Statistical Modelling in Biostatistics and Bioinformatics, Contributions to Statistics. Springer International Publishing, pp 139–157
4. Koestler DC, Marsit CJ, Christensen BC, Kelsey KT, Houseman EA (2014) A recursively partitioned mixture model for clustering time-course gene expression data. Translational Cancer Research 3(3)
5. Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V (2014) Hiv haplotype inference using a propagating dirichlet process mixture model. IEEE/ACM Trans Comput Biol Bioinform 11(1):182–191
6. Tran KA, Vo NQ, Lee G (2014) A novel clustering algorithm based gaussian mixture model for image segmentation. In: Proc. of the 8th International Conference on Ubiquitous Information Management and Communication, ICUIMC '14, pp 97:1–97:5 ACM
7. Topkaya IS, Erdogan H, Porikli F (2014) Counting people by clustering person detector outputs. In: Proc. of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 313–318
8. Zhou B, Tang X, Wang X (2015) Learning collective crowd behaviors with dynamic pedestrian-agents. Int J Comput Vis 111(1):50–68
9. Boutemedjet S, Ziou D (2012) Predictive approach for user long-term needs in content-based image suggestion. IEEE Transactions on Neural Networks and Learning Systems 23(8):1242–1253
10. Beutel A, Murray K, Faloutsos C, Smola AJ (2014) Cobafi: Collaborative bayesian filtering. In: Proc. of the 23rd International Conference on World Wide Web, WWW '14, pages 97–108. ACM
11. Yin H, Cui B, Chen L, Hu Z, Huang Z (2014) A temporal context-aware model for user behavior modeling in social media systems. In: Proc. of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14, pp 1543–1554. ACM
12. Handcock MS, Raftery AE, Tantrum JM (2007) Model-based clustering for social networks. J R Stat Soc: Series A (Statistics in Society) 170(2):301–354
13. Couronne T, Stoica A, Beuscart JS (2010) Online social network popularity evolution: An additive mixture model. In: Proc. of International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp 346–350
14. Xu D, Yang S (2014) Location prediction in social media based on contents and graphs. In: Proc. of Fourth International Conference on Communication Systems and Network Technologies (CSNT), pp 1177–1181
15. Bdiri T, Bouguila N (2011) Learning inverted dirichlet mixtures for positive data clustering . In: Proc. of the 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC), pp 265–272

16. Bdiri T, Bouguila N (2012) Positive vectors clustering using inverted dirichlet finite mixture models. Expert Systems With Applications 39(2):1869–1882
17. Bdiri T, Bouguila N, Ziou D (2014) Object clustering and recognition using multi-finite mixtures for semantic classes and hierarchy modeling. Expert Systems with Applications 41(4, Part 1):1218–1235
18. Bdiri T, Bouguila N, Ziou D (2015) A statistical framework for online learning using adjustable model selection criteria. Technical report, Concordia Institute for Information Systems Engineering. Concordia University, Montreal
19. Bdiri T, Bouguila N, Ziou D (2013) Visual scenes categorization using a flexible hierarchical mixture model supporting users ontology. In: IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI), pp 262–267
20. Wallace CS (2005) Statistical and inductive inference by minimum message length. Springer-Verlag
21. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
22. Rissanen J (1978) Modeling by shortest data description. Automatica 14(5):465–471
23. Figueiredo MAT, Leit ao JMN, Jain A (1999) On fitting mixture models. In: Proc. of the Second International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer-Verlag, pp 54–69
24. McLachlan GJ, Peel D (2000) Finite Mixture Models. Wiley, New York
25. McLachlan GJ, Krishnan T (1997) The EM Algorithm and Extensions. John Wiley and Sons. Inc.
26. Winn J, Bishop CM (2005) Variational Message Passing. J Mach Learn Res 6:661–694
27. Dimitris K, Evdokia X (2003) Choosing initial values for the {EM} algorithm for finite mixtures. Comput Stat Data Anal 41(34):577–590
28. Robert CP (2007) The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd edn. Springer
29. Bouguila N, Elguebaly T (2012) A fully bayesian model based on reversible jump {MCMC} and finite beta mixtures for clustering. Expert Systems with Applications 39(5):5946–5959
30. Pereyra M, Dobigeon N, Batatia H, Tourneret J (2013) Estimating the granularity coefficient of a potts-markov random field within a markov chain monte carlo algorithm. IEEE Trans Image Process 22(6):2385–2397
31. Bouguila N, Ziou D (2008) A dirichlet process mixture of dirichlet distributions for classification and prediction. In: IEEE Workshop on Machine Learning for Signal Processing (MLSP), pp 297–302
32. Cowles MK, Carlin BP (1996) Markov chain monte carlo convergence diagnostics: A comparative review. J Am Stat Assoc 91(434):883–904
33. Bhatnagar N, Bogdanov A, Mossel E (2011) The computational complexity of estimating mcmc convergence time. In: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, volume 6845 of Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp 424–435
34. Corduneanu A, Bishop CM (2001) Variational bayesian model selection for mixture distributions. In: Proc. of the Eighth International Conference on Artificial Intelligence and Statistics, p 2734. Morgan Kaufmann
35. Tan SL, Nott DJ (2014) Variational approximation for mixtures of linear mixed models. J Comput Graph Stat 23(2):564–585
36. Thanh MN, Wu QMJ (2014) Asymmetric mixture model with variational bayesian learning. In: Proc. of International Joint Conference on Neural Networks (IJCNN), pp 285–290
37. Zhanyu M, Leijon A (2011) Bayesian estimation of beta mixture models with variational inference. IEEE Trans Pattern Anal Mach Intell 33(11):2160–2173
38. Boutemedjet S, Bouguila N, Ziou D (2009) A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering. IEEE Trans Pattern Anal Mach Intell 31(8):1429–1443
39. Wang H, Zha H, Qin H (2007) Dirichlet aggregation: unsupervised learning towards an optimal metric for proportional data. In: Proceedings of the 24th international conference on Machine learning, pp 959–966. ACM
40. Johnson NL, Kotz S, Balakrishnan N (1995) Continuous Univariate Distributions: Vol.: 2. Wiley series in probability and mathematical statistics. Applied probability and statistics
41. Sethuraman J. (1994) A constructive definition of Dirichlet priors. Stat Sin 4:639–650
42. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer
43. Fei-Fei L, Fergus R, Perona P (2004) Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: Proc. of conference on Computer Vision and Pattern Recognition Workshop (CVPRW), pp 178–178
44. Constantinopoulos C, Titsias MK, Likas A (2006) Bayesian feature and model selection for gaussian mixture models. IEEE Trans Pattern Anal Mach Intell 28(6):1013–1018
45. Fan W, Bouguila N (2013) Variational learning of a dirichlet process of generalized dirichlet distributions for clustering, simultaneous feature selection. Pattern Recogn 46(10):2754–2769
46. Blei DM, Jordan MI (2006) Variational inference for dirichlet process mixtures. Bayesian Analysis 1(1):121–143
47. Jordan M, Ghahramani Z, Jaakkola T, Saul L (1999) An introduction to variational methods for graphical models. Mach Learn 37(2):183–233
48. Opper M, Saad D (2001) Advanced mean field methods: theory and practice. Neural Information Processing. Massachusetts Institute of Technology Press (MIT Press)
49. Bishop CM (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, New York Inc.
50. Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press
51. Ishwaran H, James LF (2001) Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc 96(453)
52. Figueiredo MAT, Jain AK (2002) Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Mach Intell 24(3):381–396
53. Law MHC, Figueiredo MAT, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. IEEE Trans Pattern Anal Mach Intell 26(9):1154–1166
54. Salter MT, Murphy TB (2012) Variational bayesian inference for the latent position cluster model for network data. Comput Stat Data Anal 57(1):661–671
55. Nasios N, Bors AG (2006) Variational learning for gaussian mixture models . IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 36(4):849–862
56. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vis 42:145–175

57. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 886–893 IEEE
58. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc. of the IEEE 86(11):2278–2324

Taoufik Bdiri received the computer engineer degree from the National Engineering School of Sousse, University of Sousse, Tunisia in 2008, where he was awarded valedictorian. He received the M.A.Sc degree in Quality Systems Engineering and the Ph.D. degree in Electrical and Computer Engineering, all from Concordia University, Montreal, Quebec, Canada, in 2010 and 2015 respectively. His research interests include statistical modeling, pattern recognition, machine learning, search engines, artificial intelligence and computer vision.

Nizar Bouguila received the engineer degree from the University of Tunis, Tunis, Tunisia, in 2000, and the M.Sc. and Ph.D. degrees in computer science from Sherbrooke University, Sherbrooke, QC, Canada, in 2002 and 2006, respectively. He is currently an Associate Professor with the Concordia Institute for Information Systems Engineering (CIISE) at Concordia University, Montreal, Qc, Canada. His research interests include image processing, machine learning, data mining, 3D graphics, computer vision, and pattern recognition. Prof. Bouguila received the best Ph.D Thesis Award in Engineering and Natural Sciences from Sherbrooke University in 2007. He was awarded the prestigious Prix d'excellence de l'association des doyens des études supèrieures au Québec (best Ph.D Thesis Award in Engineering and Natural Sciences in Québec), and was a runner-up for the prestigious NSERC doctoral prize. He is an IEEE senior member and the holder of a Concordia Research Chair.



Djemel Ziou received the B.Eng. degree in Computer Science from the University of Annaba (Algeria) in 1984, and Ph.D degree in Computer Science from the Institut National Polytechnique de Lorraine (INPL), France in 1991. From 1987 to 1993 he served as lecturer in several universities in France. During the same period, he was a researcher in the Centre de Recherche en Informatique de Nancy (CRIN) and the Institut National de Recherche en Informatique et Automatique (INRIA) in France. Presently, he is full Professor at the department of computer science, Université de Sherbrooke, QC, Canada. He is holder of the NSERC/Bell Canada Research Chair in personal imaging. He has served on numerous conference committees as member or chair. He heads the laboratory MOIVRE and the consortium CoRIMedia which he founded. His research interests include image processing, information retrieval, computer vision and pattern recognition.