Stat Comput (2006) 16:215–225
DOI 10.1007/s11222-006-8451-7

Practical Bayesian estimation of a finite beta mixture through gibbs sampling and its applications

Nizar Bouguila · Djemel Ziou · Ernest Monga

Received: September 2004 / Accepted: January 2006

Abstract This paper deals with a Bayesian analysis of a finite Beta mixture model. We present approximation method to evaluate the posterior distribution and Bayes estimators by Gibbs sampling, relying on the missing data structure of the mixture model. Experimental results concern contextual and non-contextual evaluations. The non-contextual evaluation is based on synthetic histograms, while the contextual one model the class-conditional densities of pattern-recognition data sets. The Beta mixture is also applied to estimate the parameters of SAR images histograms.

Keywords Beta distribution · Mixture modeling · Maximum likelihood · Bayesian analysis · Gibbs sampling · Metropolis-Hastings · EM · SEM · SAR images

1. Introduction

Finite mixtures are a flexible and powerful probabilistic tool for modeling univariate and multivariate data (McLachlan and Peel 2000, Everitt and Hand 1981). The usefulness of mixture models is currently widely acknowledged in any area which involves the statistical modeling of data such as astronomy, ecology, bioinformatics, pattern recognition, computer vision and machine learning. Mixture modeling can be viewed as the superimposition of a finite number of component densities. The Gaussian density is widely accepted and used in mixture modeling (Roberts and Rezek 1998). At the same time, other models such as Beta mixtures have not received much attention. If the random variable X follows a Beta distribution, the density function is given by Samuel et al. (1990):

$$p(X) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} X^{\alpha-1}(1-X)^{\beta-1} \tag{1}$$

where $0 < X < 1, \alpha > 0$ and $\beta > 0$. The mean and the variance of the Beta distribution are given by:

$$E(X) = \frac{\alpha}{\alpha+\beta} \tag{2}$$

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{3}$$

This distribution is the univariate case of the Dirichlet distribution which has proven to have high flexibility to model data (Bouguila et al. 2004). Indeed, the capacity of an univariate distribution to provide an accurate fit to data depends on its shape. The shape can be defined by the third β_1 and fourth β_2 moments and they represent the coefficients of the asymmetry and flatness of a given distribution. In the plane (β_1, β_2), the shape of the Gaussian distribution is represented by a point (0, 3), that of the Gamma distribution is a line, that of the Beta distribution is a plane and that of the Log-Normal is a line (see Figure 1). Then, the shapes of the Beta are variable enough to allow for an approximation of almost any arbitrary distribution. In fact, the Beta distribution permits multiple symmetric and asymmetric modes (see Figure 2). We note that the Beta distribution is defined in the compact support [0, 1] in contrast of the Gaussian, for example, which is defined in $\mathbb{R}$. However, we can generalize it easily to be defined in a compact support of the form $[A, B]$, where $(A, B) \in \mathbb{R}^2$ (Beckman and Tietjen 1978). Having a compact support is an interesting property for a given density because of the nature of data in general. Generally, we estimate data which are compactly supported, such as data

N. Bouguila · D. Ziou (✉) · E. Monga
Département d'Informatique, Faculté des Sciences,
Université de Sherbrooke, Sherbrooke, Qc,
Canada J1K 2R1
e-mail: {nizar.bouguila, djemel.ziou, ernest.monga}@usherbrooke.ca

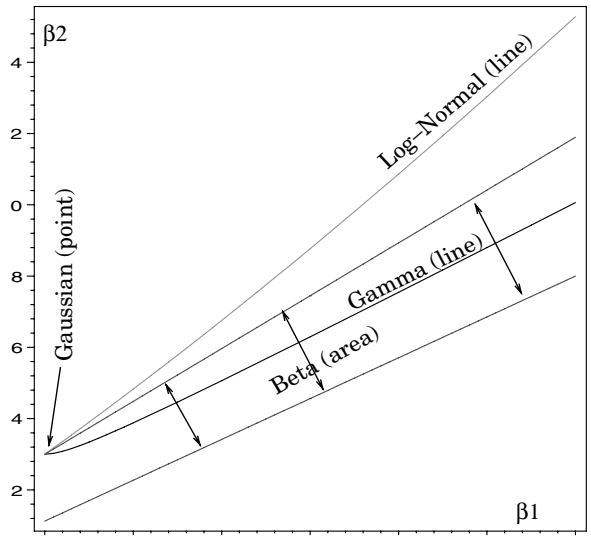

Fig. 1 Representation of the Gaussian, Gamma, Beta and Log-Normal distributions in the (β_1, β_2) plane

originating from videos, images or text. The Beta distribution has already been used by many authors by playing an auxiliary role as a prior to the binomial which was the parent distribution (Klieter 1992, Lee and Lio 1999). In this paper, we assume that the parent is Beta[1].

The estimation of the parameters of finite mixture distribution has recently come under close scrutiny (McLachlan and Peel 2000). Computational methods, have also appeared, including the EM algorithm proposed by Dempster et al. (1977). However, the EM algorithm for finite mixture has several drawbacks (McLachlan and Krishnan 1997). These drawbacks are mainly optimization problems. For example, the occurrence of local modes of a likelihood will often cause problems for a deterministic gradient method (Robert and Casella 1999). In recent, developments of computational methods, Bayesian methods are considered as an alternative way to deal with mixture models. Given a proper prior, a Bayesian approach to the mixture estimation problem always provides estimators which can be written explicitly for conjugate priors (Diebolt and Robert 1994). Besides, Bayesian approaches are based on simulation methods, such as Gibbs sampling, which explore high-density regions. Diebolt and Robert (1994) used data augmentation and Gibbs sampling as approximation methods for evaluating the posterior distribution and Bayes estimators for Gaussian mixture. Tsung et al. (2004) used Bayesian inference for finite t distribution. Tsionas (2004) considers the estimation of the parameters of the multivariate Gamma distribution using Gibbs sampling with data augmentation. In this article, we deal with a mixture of Beta distributions form Bayesian viewpoints. The choice of the Beta distribution is justified by the interesting properties of this distribution that we have explained above. It is noted that this paper deals with the issue of estimating the parameters of a finite Beta mixture assuming that the sample corresponds exactly to a M-components. The paper is organized as follows. The next section describes the Beta mixture and the Bayesian approach in details. The complete estimation algorithm is given, too. Section 3 is devoted to experimental results. We end the paper with some concluding remarks.

2. The finite beta mixture and Bayesian estimtion

2.1. The model

A Beta mixture with M components is defined as:

$$p(X|\Theta) = \sum_{j=1}^{M} p(X|\xi_j)P_j \tag{4}$$

where P_j $(0 < P_j \leq 1$ and $\sum_{j=1}^{M} P_j = 1)$ are the mixing proportions and $p(X|\xi_j)$ is the Beta distribution. The symbol $\Theta = (\xi, P)$ refers to the entire set of parameters to be estimated, where $\xi = (\xi_1, \ldots, \xi_M)$, $\xi_j = (\alpha_j, \beta_j)$ is the parameter vector for the j^{th} population and $P = (P_1, \ldots, P_M)$. Consider N independent observations $\mathcal{X} = (X_1, \ldots, X_N)$, the likelihood corresponding to a M-component is:

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} \sum_{j=1}^{M} p(X_i|\xi_j)P_j \tag{5}$$

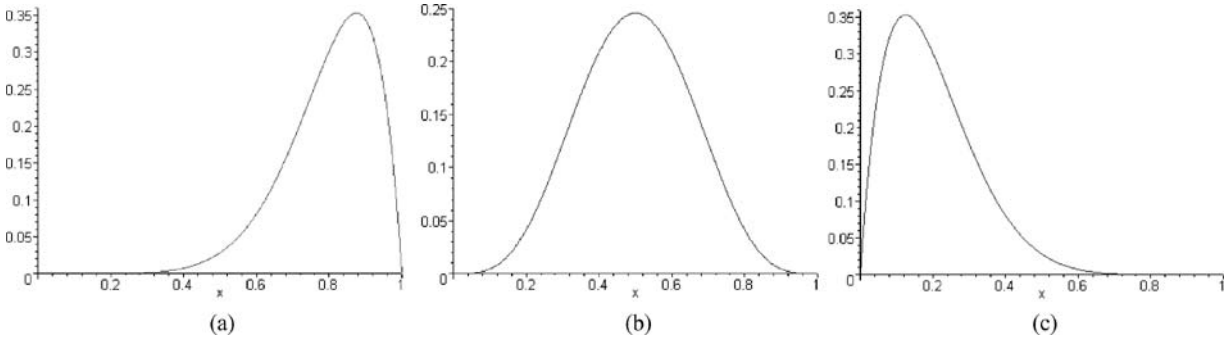

Fig. 2 The Beta distribution for different parameters. (a) A Beta distribution skewed to the right: $\alpha = 8, \beta = 2$ (b) A symmetric Beta distribution: $\alpha = 5, \beta = 5$ (c) A Beta distribution skewed to the left: $\alpha = 2, \beta = 8$

For maximum likelihood computations, it's possible to use numerical optimization procedure like EM algorithm (Dempster et al. 1977). With the EM algorithm, the mixture model is expressed in terms of missing data. If, for each variable X_i, $1 < i < N$, $Z_i = (Z_{i1}, \ldots, Z_{iM})$ is a M-dimensional vector indicating to which component X_i belongs, such that:

$$Z_{ij} = \begin{cases} 1 & \text{if } X_i \text{ belongs to class } j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The complete-data likelihood is then:

$$p(\mathcal{X}, \mathcal{Z}|\Theta) = \prod_{i=1}^{N} \prod_{j=1}^{M} (P_j p(X_i|\xi_j))^{Z_{ij}} \tag{7}$$

and the complete-data loglikelihood is:

$$L(\Theta, \mathcal{Z}, \mathcal{X}) = \sum_{j=1}^{M} \sum_{i=1}^{N} Z_{ij} log(P_j p(X_i|\xi_j)) \tag{8}$$

where $\mathcal{Z} = \{Z_1, \ldots, Z_N\}$. From this perspective, the EM algorithm can be used to estimate the mixture parameters, replacing each missing data Z_{ij} by its expectation (McLachlan and Peel 2000):

$$\hat{Z}_{ij}^{(t)} = \frac{p^{(t-1)}\left(X_i|\xi_j^{(t-1)}\right) P_j^{(t-1)}}{\sum_{j=1}^{M} p^{(t-1)}\left(X_i|\xi_j^{(t-1)}\right) P_j^{(t-1)}} \tag{9}$$

where t indexes the current iteration step and $P_j^{(t)}$ and $\xi_j^{(t)}$ are the current evaluation of the parameters. The EM algorithm produces a sequence of estimate $\{\Theta^{(t)}, t = 0, 1, 2, \ldots\}$ by alternately applying two steps (until some convergence criterion is satisfied):

1. E-step: Compute $\hat{Z}_{ij}^{(t)}$ (Eq. 9) parameter estimates from the initialization.
2. M-step: Update the parameters estimates according to:

$$\hat{\Theta}^{(t)} = argmax_{\Theta} L(\Theta^{(t-1)}, \mathcal{Z}, \mathcal{X})$$

where $\hat{Z}_{ij}^{(t)}$ is the posterior probability that the ith observation arises form the jth component of the mixture. The EM algorithm has some disadvantages. In fact, problems with the EM algorithm can occur in the case of multimodal likelihoods. The increase of the likelihood function at each step of the algorithm ensures its convergence to the maximum likelihood estimator in the case of unimodal likelihoods but implies a dependance on initial conditions for multimodal likelihoods. In this last case, it happens that the EM converges to a saddle point but not to a local maximum. Several extensions to the EM algorithm can be found in the literature to overcome these problems. A lot of these algorithms are based on Bayesian approaches.

2.2. Bayesian estimation

In order to overcome the problems of numerical methods presented above, simulation methods are often chosen as a solution. In general, these methods are related to the Bayesian theory. In the Bayesian paradigm information brought by the complete data $(\mathcal{X}, \mathcal{Z})$, a realization of $(\mathcal{X}, \mathcal{Z}) \sim p(\mathcal{X}, \mathcal{Z}|\Theta)$, is combined with prior information about the parameters Θ that is specified in a prior distribution with density $\pi(\Theta)$ and summarized in a probability distribution $\pi(\Theta|\mathcal{X}, \mathcal{Z})$, called the posterior distribution (Robert and Casella 1999). This is derived from the joint distribution $p(\mathcal{X}, \mathcal{Z}|\Theta)\pi(\Theta)$, according to Bayes formula:

$$\pi(\Theta|\mathcal{X}, \mathcal{Z}) = \frac{p(\mathcal{X}, \mathcal{Z}|\Theta)\pi(\Theta)}{\int p(\mathcal{X}, \mathcal{Z}|\Theta)\pi(\Theta)} \propto p(\mathcal{X}, \mathcal{Z}|\Theta)\pi(\Theta) \tag{10}$$

Having this posterior distribution in hand we can simulate $\Theta \sim \pi(\Theta|\mathcal{X}, \mathcal{Z})$ rather than computing them. This simulation technique is now well known as Gibbs sampling. However, prior to the appearance of the Gibbs sampling algorithm, Celeux and Diebolt considered a modified version of the EM algorithm in the context of computing the maximum likelihood estimators for finite mixture models (Celeux and Diebolt, 1985). They call it the Stochastic EM algorithm which we can connect to the Gibbs sampling technique. These two techniques and their connection are discussed in what follows.

One of the most successful extension of the EM was the SEM algorithm that was elaborated by Celeux and Diebolt (1985). The SEM algorithm consists in fact of a modification EM algorithm in which a probabilistic teacher step (Stochastic step or S-Step) has been incorporated. This step can be viewed as a Bayesian extension of the EM algorithm. In fact, it consists of a simulation of $\mathcal{Z}$ according to the posterior probability $\pi(\mathcal{Z}|\mathcal{X}, \Theta)$. This posterior probability is chosen to be Multinomial of order one with weights given by the $\hat{Z}_{ij}$ ($\mathcal{M}(1; \hat{Z}_{i1}, \ldots, \hat{Z}_{iM})$). This choice can be explained intuitively. Indeed, we know that each Z_i is a vector of zero-one indicator variables that define the component from which the ith observation X_i arises. Besides, the probability that the ith observation arises form the jth component of the mixture is given by $\hat{Z}_{ij}$. So, we can think that each vector Z_i is generated by a Multinomial distribution of order one with weights given by the $\hat{Z}_{ij}$. The SEM algorithm is summed up as follows:

1. E-step: Compute $\hat{Z}_{ij}^{(t)}$ parameter estimates from the initialization.
2. S-step: For each observation X_i, $i = 1, \ldots, N$, a draw $Z_i^{(t)} = (Z_{i1}^{(t)}, \ldots, Z_{iM}^{(t)})$ is made from the multinomial distribution with probabilities $(\hat{Z}_{i1}^{(t)}, \ldots, \hat{Z}_{iM}^{(t)})$. Notice that

exactly one of these $Z_{ij}^{(t)}$, $j = 1, \ldots, M$ is 1 and the others 0 for each j.

3. M-step: Update the parameters estimates according to:

$$\hat{\Theta}^{(t)} = argmax_{\Theta} L(\Theta^{(t-1)}, \mathcal{Z}, \mathcal{X})$$

Celeux and Diebolt have shown that the S-step prevents the estimates from staying near saddle point of the likelihood function (Celeux and Diebolt 1992). They give a mathematical proof, based on Markov chains, to show the convergence of this algorithm, too Celeux and Diebolt (1985). The appeal of this approach is that it allows for a more systematic exploration of the likelihood surface by partially avoiding the fatal attraction of the closest mode.

Notice that the calculation of the M-step depends upon the form of the density. Often in practice, the solution of this step does not exist in closed form. In this situation, consideration may be given to using some iterative schemes. However, another problem occurs. In fact, the parameters of the Beta can become very high during iterations (see the first example in section 3) and causes numeric problems. An efficient solution can be simulating the vector of parameters Θ, rather than computing it. This approach is in fact a logical extension of the SEM algorithm, which complete the corresponding simulation of $\mathcal{Z}$ and which is now well know as Gibbs sampling.

The Gibbs sampler is the most commonly used approach in Bayesian mixture estimation (Escobar and West 1995, Diebolt and Robert 1994). In fact, a solution to the computational problem is to take advantage of the missing data introduced in the previous section, that is to associate with each observation X_i a missing multinomial variable $Z_i \sim \mathcal{M}(1; \hat{Z}_{i1}, \ldots, \hat{Z}_{iM})$. Recall that $\mathcal{Z} = \{Z_1, \ldots, Z_N\}$ and denote by $\pi(P|\mathcal{Z}, \mathcal{X})$ the density of the distribution of P given $\mathcal{Z}$ and $\mathcal{X}$. This distribution is in fact independent of $\mathcal{X}$, $\pi(P|\mathcal{Z}, \mathcal{X}) = \pi(P|\mathcal{Z})$ (Marin et al. 2004). The standard Gibbs sampler for mixture models is based on the successive simulation of $\mathcal{Z}$, P and ξ (Diebolt and Robert 1994, Marin et al. 2004):

1. Initialization
2. Step t: For $t = 1, \ldots$
   (a) Generate $Z_i^{(t)} \sim \mathcal{M}(1; \hat{Z}_{i1}^{(t-1)}, \ldots, \hat{Z}_{iM}^{(t-1)})$
   (b) Generate P from $\pi(P|\mathcal{Z}^{(t)})$
   (c) Generate ξ from $\pi(\xi|\mathcal{Z}^{(t)}, \mathcal{X})$

We start by the distribution $\pi(P|\mathcal{Z}, \mathcal{X})$ and we have:

$$\pi(P|\mathcal{Z}) \quad \propto \quad \pi(P)\pi(\mathcal{Z}|P) \tag{11}$$

We determine now $\pi(P)$ and $\pi(\mathcal{Z}|P)$. We know that the vector P is defined on the simplex $\{(P_1, \ldots, P_M) : \sum_{j=1}^{M-1} P_j < 1\}$, then a natural choice, as a prior, for this vector is the Dirichlet distribution (Marin et al. 2004):

$$\pi(P) = \frac{\Gamma\left(\sum_{j=1}^{M} \eta_j\right)}{\prod_{j=1}^{M} \Gamma(\eta_j)} \prod_{j=1}^{M} P_j^{\eta_j - 1} \tag{12}$$

where $\eta = (\eta_1, \ldots, \eta_M)$ is the parameter vector of the Dirichlet distribution. Moreover, we have:

$$\begin{aligned} \pi(\mathcal{Z}|P) &= \prod_{i=1}^{N} \pi(Z_i|P) = \prod_{i=1}^{N} P_1^{Z_{i1}} \ldots P_M^{Z_{iM}} \\ &= \prod_{i=1}^{N} \prod_{j=1}^{M} P_j^{Z_{ij}} = \prod_{j=1}^{M} P_j^{n_j} \end{aligned} \tag{13}$$

where $n_j = \sum_{i=1}^{N} \mathbb{I}_{Z_{ij}=j}$. Then

$$\begin{aligned} \pi(P \mid \mathcal{Z}) &= \frac{\Gamma\left(\sum_{j=1}^{M} \eta_j\right)}{\prod_{j=1}^{M} \Gamma(\eta_j)} \prod_{j=1}^{M} P_j^{\eta_j - 1} \prod_{j=1}^{M} P_j^{n_j} \\ &= \frac{\Gamma\left(\sum_{j=1}^{M} \eta_j\right)}{\prod_{j=1}^{M} \Gamma(\eta_j)} \prod_{j=1}^{M} P_j^{\eta_j + n_j - 1} \\ &\propto \mathcal{D}(\eta_1 + n_1, \ldots, \eta_M + n_M) \end{aligned} \tag{14}$$

where $\mathcal{D}$ is a Dirichlet distribution with parameters $(\eta_1 + n_1, \ldots, \eta_M + n_M)$. We note that the prior and the posterior distributions, $\pi(P)$ and $\pi(P|\mathcal{Z})$, are both Dirichlet. In this case we say that the Dirichlet distribution is a conjugate prior for the mixture proportions.

For a mixture of Beta distributions, it is therefore possible to associate with each $\xi_j = (\alpha_j, \beta_j)$ a prior $\pi_j(\xi_j)$. The $\xi_j = (\alpha_j, \beta_j)$ parameters of the Beta can be understood by considering the following alternative representation:

$$s_j = \alpha_j + \beta_j \tag{15}$$

$$m_j = \alpha_j / s_j \tag{16}$$

and then equation (1) can be written as follows:

$$\begin{aligned} p(X_i|s_j, m_j) \; &= \frac{\Gamma(s_j)}{\Gamma(s_j m_j)\Gamma(s_j(1 - m_j))} \\ &\quad X_i^{s_j m_j}(1 - X_i)^{s_j(1 - m_j)} \end{aligned} \tag{17}$$

Based on this representation, we use a prior previously proposed by Robert and Rousseau (2002):

$$\begin{aligned} \pi(s_j, m_j) \propto (1 - &exp\,(-\delta((s_j - 2)^2 + (m_j - 0.5)^2))) \\ &exp\left(\frac{-\rho}{s_j^2 m_j(1 - m_j)} - \kappa s_j^2/2)\right) \end{aligned} \tag{18}$$

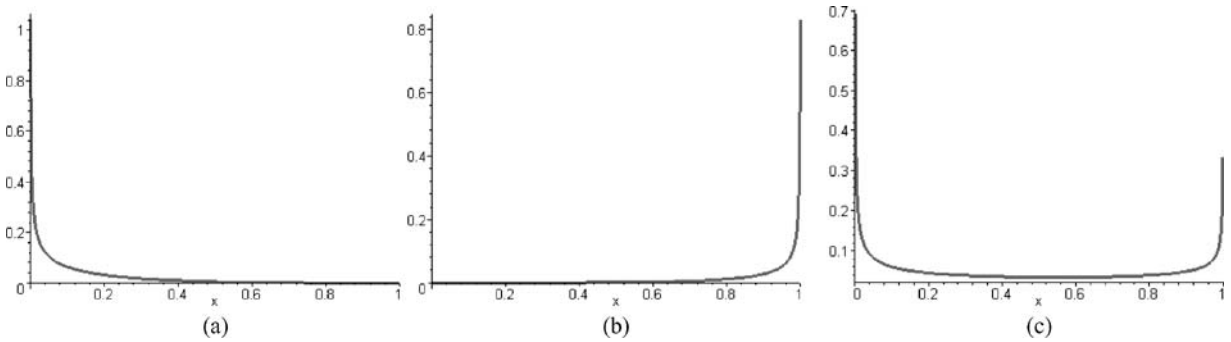

Fig. 3 Examples of histograms generated by small values of α_j and β_j. (a) $\alpha_j = 0.5$ and $\beta_j = 4$. (b) $\alpha_j = 5$ and $\beta_j = 0.4$. (c) $\alpha_j = 0.5$ and $\beta_j = 0.6$

where δ, ρ and κ are hyperparameters. There surely a fair amount of arbitrariness with this choice, but it fits our goals. Indeed, this choice is designed to avoid the $(s_j, m_j) = (2, 0.5)$. In this last case, we obtain $(\alpha_j, \beta_j) = (1, 1)$ and then the Beta density will be proportional to 1 all the time (constant histograms). In addition, by this choice we exclude the small values of α_j and β_j ($\alpha_j < 1$ or $\beta_j < 1$) which are not of interest (See Figure 3). More details about this prior are given in (Robert and Rousseau 2002). Having this prior, $\pi(s_j, m_j)$, the posterior distribution is then:

$$\pi((s_j, m_j) \mid \mathcal{Z}, \mathcal{X}) \propto \pi(s_j, m_j) \prod_{Z_{ij}=1} p(X_i \mid s_j, m_j) \tag{19}$$

$$\begin{aligned} &\propto (1 - exp(-\delta((s_j - 2)^2 + (m_j - 0.5)^2))) \\ &\times exp\left(\frac{-\rho}{s_j^2 m_j(1 - m_j)} - \kappa s_j^2/2)\right) \\ &\times \left(\frac{\Gamma(s_j)}{\Gamma(s_j m_j)\Gamma(s_j(1 - m_j))}\right)^{n_j} \left(\prod_{Z_{ij}=1} X_i\right)^{s_j m_j} \\ &\times \left(\prod_{Z_{ij}=1} (1 - X_i)\right)^{s_j(1-m_j)} \end{aligned} \tag{20}$$

2.3. Algorithm

Having all the posterior probabilities in hand, the steps of the Gibbs sampler are:

1. Initialization
2. Step t: For $t = 1, \ldots$
   (a) Generate $Z_i^{(t)} \sim \mathcal{M}(1; \hat{Z}_{i1}^{(t-1)}, \ldots, \hat{Z}_{iM}^{(t-1)})$
   (b) Compute $n_j^{(t)} = \sum_{i=1}^{N} \mathbb{I}_{Z_{ij}^{(t)}=j}$
   (c) Generate $P^{(t)}$ from Eq. 14
   (d) Generate $(s_j, m_j)^{(t)}$ $(j = 1, \ldots, M)$ from Eq. 19 using the Metropolis-Hastings (M-H) algorithm[2].

The M-H algorithm (Robert and Casella 1999) offers a solution to the problem of simulating from the posterior distribution. Starting from point $(s_j^{(0)}, m_j^{(0)})$, the corresponding Markov chain explores the surface of the posterior distribution. At iteration t, the steps of the M-H algorithm can be described as follows:

1. Generate $(\tilde{s_j}, \tilde{m_j}) \sim q(s_j, m_j | s_j^{(t-1)}, m_j^{(t-1)})$ and $U \sim \mathcal{U}_{[0,1]}$
2. Compute
$$r = \frac{\pi(\tilde{s_j}, \tilde{m_j} | \mathcal{Z}, \mathcal{X}) q(s_j^{(t-1)}, m_j^{(t-1)} | \tilde{s_j}, \tilde{m_j})}{\pi(s_j^{(t-1)}, m_j^{(t-1)} | \mathcal{Z}, \mathcal{X}) q(\tilde{s_j}, \tilde{m_j} | s_j^{(t-1)}, m_j^{(t-1)})}$$
3. If $r < u$ then $(s_j^{(t)}, m_j^{(t)}) = (\tilde{s_j}, \tilde{m_j})$ else $(s_j^{(t)}, m_j^{(t)}) = (s_j^{(t-1)}, m_j^{(t-1)})$

The major problem in this algorithm is the need to choose the proposal distribution q. The most generic proposal is the random walk Metropolis-Hastings algorithm where each unconstrained parameter is the mean of the proposal distribution for the new value. As $\tilde{s_j} > 0$, we have chosen the following proposal:

$$\tilde{s_j} \sim \mathcal{LN}\left(log\left(s_j^{(t-1)}\right), \sigma^2\right) \tag{21}$$

where $\mathcal{LN}(log(s_j^{(t-1)}), \sigma^2)$ refers to the log-normal distribution with mean $log(s_j^{(t-1)})$ and variance σ^2. Note that equation (21) is equivalent to:

$$log(\tilde{s_j}) = log\left(s_j^{(t-1)}\right) + \epsilon_j \tag{22}$$

where $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$. However, for constrained parameters, this proposal is not efficient (Casella et al. 2000). This is the case for the parameter m_j (since m_j belongs to the simplex [0, 1]). To resolve this difficulty, we first transform m_j, $j = 1, \ldots, M$ to $m_j^* = m_j/(1 - m_j)$ and then use the following proposal:

$$\tilde{m_j^*} \sim \mathcal{LN}(log(m_j^{*(t-1)}), \sigma^2) \tag{23}$$

with these proposals the random walk M-H algorithm is composed of the following steps:

1. Generate $\tilde{s_j} \sim \mathcal{LN}(log(s_j^{(t-1)}), \sigma^2)$, $\tilde{m_j^*} \sim \mathcal{LN}(log(m_j^{*(t-1)}), \sigma^2)$ and $U \sim \mathcal{U}_{[0,1]}$.
2. Compute $r = \frac{\pi(\tilde{s_j}, \tilde{m_j^*}|\mathcal{Z},\mathcal{X})\mathcal{LN}(m_j^{*(t-1)}|log(\tilde{m_j^*}),\sigma^2)\mathcal{LN}(s_j^{(t-1)}|log(\tilde{s}_j),\sigma^2)}{\pi(s_j^{(t-1)}, m_j^{*(t-1)}|\mathcal{Z},\mathcal{X})\mathcal{LN}(\tilde{m_j^*}|log(m_j^{*(t-1)}),\sigma^2)\mathcal{LN}(\tilde{s}_j|log(s_j^{(t-1)}),\sigma^2)}$
3. If $r < u$ then $(s_j^{(t)}, m_j^{*(t)}) = (\tilde{s_j}, \tilde{m_j^*})$ else $(s_j^{(t)}, m_j^{*(t)}) = (s_j^{(t-1)}, m_j^{*(t-1)})$

where

$$\pi\left((s_j, m_j^*)|\mathcal{Z}, \mathcal{X}\right) \propto (1 - exp(-\delta((s_j - 2)^2 + (T(m_j^*) - 0.5)^2)))exp\left(\frac{-\rho}{s_j^2 T(m_j^*)(1 - T(m_j^*))} - \kappa s_j^2/2)\right) \times \left(\frac{\Gamma(s_j)}{\Gamma(s_j T(m_j^*))\Gamma(s_j(1 - T(m_j^*)))}\right)^{n_j} \left(\prod_{Z_{ij}=1} X_i\right)^{s_j T(m_j^*)} \times \left(\prod_{Z_{ij}=1} (1 - X_i)\right)^{s_j(1-T(m_j^*))} J(m_j^*)$$

where $T(m_j^*) = e^{m_j^*}/(1 + e^{m_j^*})$ and $J(m_j^*) = e^{m_j^*}/(1 + e^{m_j^*})^2$ represents the Jacobian of the transformation $T(m_j^*)$.

For the initialization, we have used the Fuzzy C-means (Bezdek 1981) and the method of moments (MM) (Fielitz and Myers 1975). In fact, the method of moments gives really good initial estimates because of the compact support of the Beta distribution. Thus, our initialization method can be summed up as follows:

Initialization algorithm

1. Apply the Fuzzy C-means to obtain the elements, variance and mean of each component.
2. Apply the MM for each component j to obtain the vector of parameters ξ_j.
3. Assign the data to clusters, assuming that the current model is correct.
4. Update the P_j using this equation:

$$P_j = \frac{\text{Number of elements in class } j}{N} \tag{24}$$

5. If the current model and the new model are sufficiently close to each other, terminate, else go to 2.

3. Experimental results

In this section, we validate the Beta mixture using contextual and non-contextual evaluations to test the performance of our method. In these applications our specific choice for the hyperparameters is $(\delta, \rho, \kappa) = (3, 0.1, 0.00001)$, which corresponds to Figure 4. This choice is designed to give the importance to the most realistic values of (α_j, β_j). For the M-H algorithm we have chosen $\sigma^2 = 0.01$. The non-contextual evaluation concerns the estimation of artificial histograms and present a case where the maximum likelihood estimation of the Beta mixture does not work. The goal of the non-contextual evaluation is to show that the estimates produced by our algorithm are accurate. The contextual evaluation is based on pattern recognition and image processing applications. The goal of the contextual evaluation is to compare the modeling capabilities of the Beta and Gaussian mixtures.

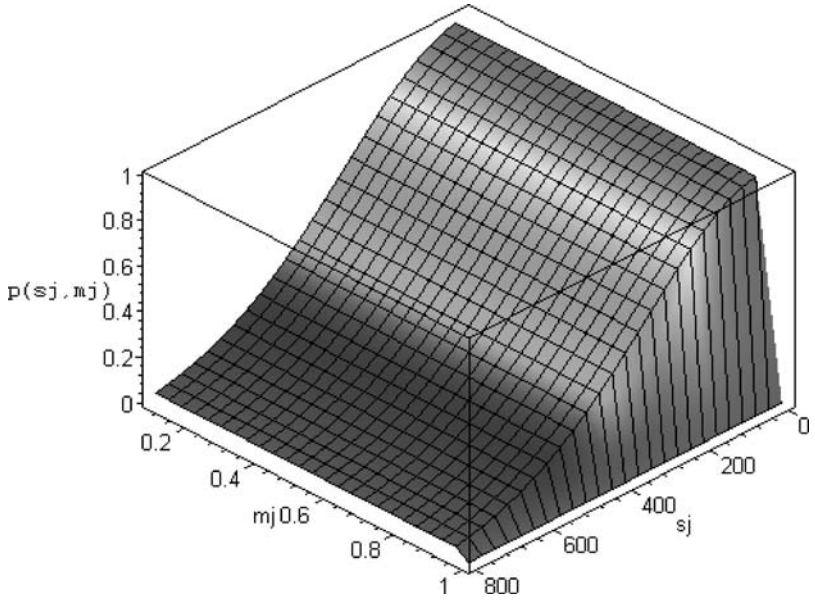

Fig. 4 Representation of the prior distribution for $(\delta, \rho, \kappa) = (3, 0.1, 0.00001)$

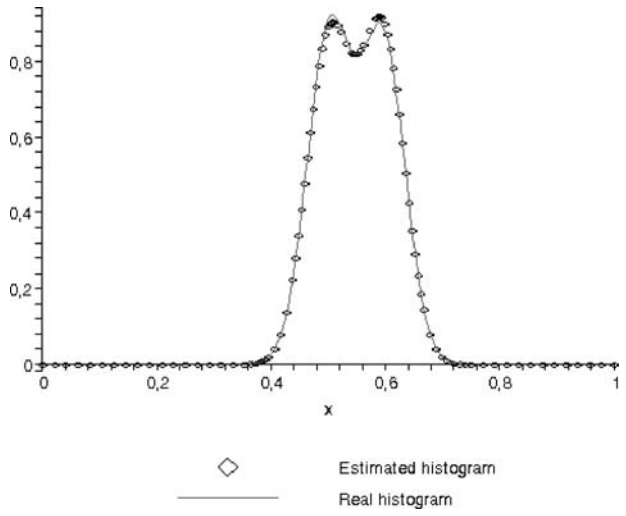

Fig. 5 First artificial histogram

3.1. Artificial histograms

We begin with the non-contextual evaluation. For this purpose, we generate artificial histograms from artificial Beta mixture models using this equation:

$$H(X_i) = \sum_{j=1}^{M} P_j \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} X_i^{\alpha_j - 1}(1 - X_i)^{\beta_j - 1} \tag{25}$$

where $i = 1 \ldots N$, N is the number of data used to generate the histogram. After that, we estimated the parameters

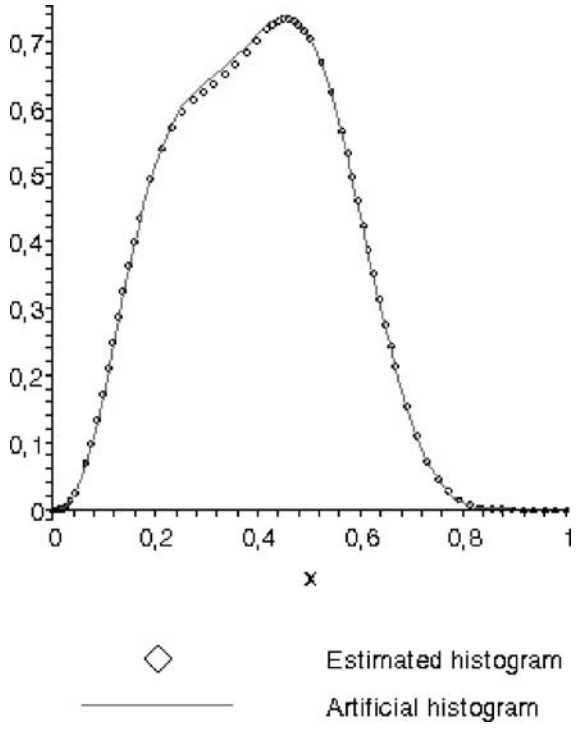

Fig. 7 Third artificial histogram

of these artificial histograms. Figures 5, 6, 7 and 8 are examples of these histograms. The first histogram (see Figure 5) presents a Beta mixture of two components. This histogram seems easy to estimate, yet we have got problems

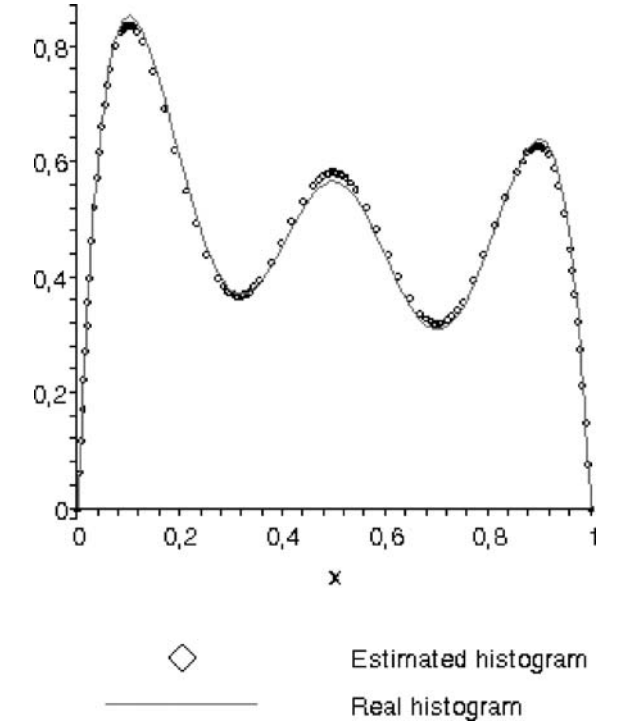

Fig. 6 Second artificial histogram

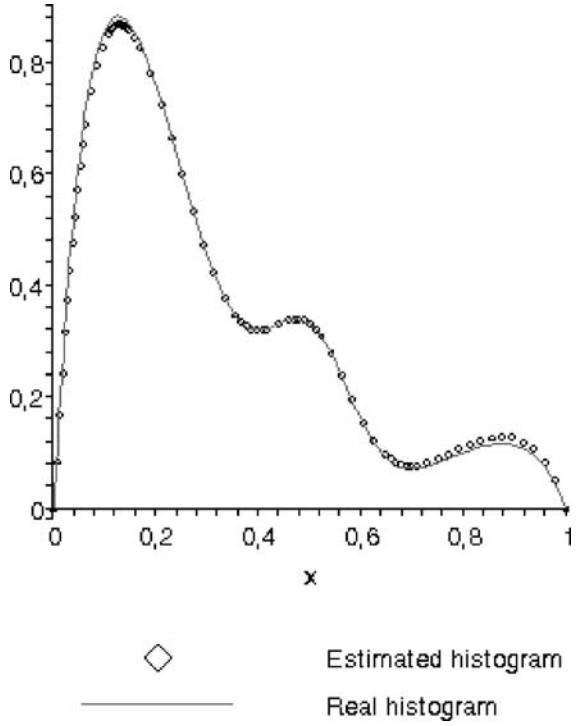

Fig. 8 Fourth artificial histogram

Table 1 Estimation of the parameters of the first artificial histogram

	Real parameters	Estimated parameters
Mode 1	$P(1) = 0.50$	$P(1) = 0.49$
	$s_1 = 170$	$s_1 = 169.06$
	$m_1 = 0.50$	$m_1 = 0.49$
Mode 2	$P(2) = 0.50$	$P(2) = 0.51$
	$s_2 = 160$	$s_2 = 160.28$
	$m_2 = 0.60$	$m_2 = 0.59$

with the ML method. The values of parameters become very high, the value of the likelihood function has gone to 0 and then the execution of other iterations was impossible. Table 1 presents the real and the estimated parameters of this histogram (when we use Bayesian estimation). The second histogram presents a Beta mixture of three well separated components. The third and fourth histograms present overlapped Beta components. The real and estimated parameters of these histograms are specified in Table 2.

Table 2 Estimation of the parameters of the second, third and fourth artificial histograms

	Real parameters	Estimated parameters
	Histogram 2	
Mode 1	$P(1) = 0.30$	$P(1) = 0.31$
	$s_1 = 20$	$s_1 = 19.86$
	$m_1 = 0.50$	$m_1 = 0.50$
Mode 2	$P(2) = 0.30$	$P(2) = 0.30$
	$s_2 = 12$	$s_2 = 11.94$
	$m_2 = 0.83$	$m_2 = 0.82$
Mode 3	$P(3) = 0.40$	$P(3) = 0.39$
	$s_3 = 12$	$s_3 = 11.97$
	$m_3 = 0.16$	$m_3 = 0.16$
	Histogram 3	
Mode 1	$P(1) = 0.50$	$P(1) = 0.49$
	$s_1 = 14$	$s_1 = 13.97$
	$m_1 = 0.28$	$m_1 = 0.29$
Mode 2	$P(2) = 0.50$	$P(2) = 0.51$
	$s_2 = 20$	$s_2 = 19.84$
	$m_2 = 0.50$	$m_2 = 0.50$
	Histogram 4	
Mode 1	$P(1) = 0.75$	$P(1) = 0.74$
	$s_1 = 10$	$s_1 = 10.08$
	$m_1 = 0.20$	$m_1 = 0.19$
Mode 2	$P(2) = 0.15$	$P(2) = 0.15$
	$m_2 = 40$	$m_2 = 39.91$
	$s_2 = 0.50$	$s_2 = 0.49$
Mode 3	$P(3) = 0.10$	$P(3) = 0.11$
	$s_3 = 10$	$s_3 = 9.95$
	$m_3 = 0.80$	$m_3 = 0.81$

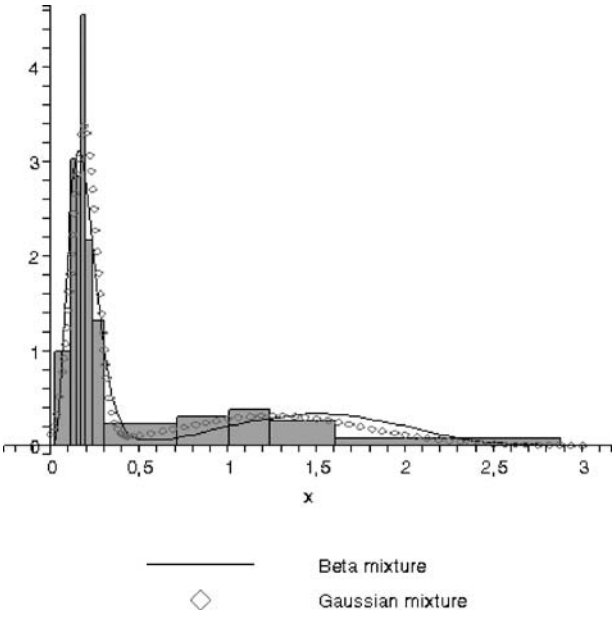

Fig. 9 Real and estimated histograms for the Enzyme data set

3.2. Real data

In the contextual applications, we validate the Beta mixture by pattern recognition and image processing applications. In the pattern recognition application, our method was used to model the class-conditional densities in 2 standard pattern recognition data sets. The first data set describes an enzymatic activity distribution in the blood among a group of 245 unrelated individuals and the second one an acidity index distribution for 155 lakes. For these two data sets, a mixture of 2 distributions is identified (Crawford 1994). Figures 9 and 10 show the real and the estimated histograms, using both Beta

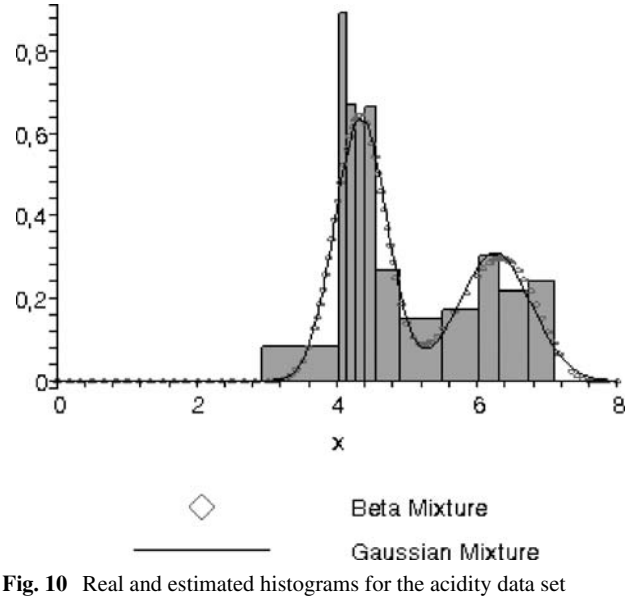

Fig. 10 Real and estimated histograms for the acidity data set

Table 3 Estimation of the parameters of the Beta and Gaussian mixtures for the Enzyme data set

	Beta mixture			Gaussian mixture		
Parameters	P_j	s_j	m_j	P_j	μ_j	σ_j^2
Mode 1	0.61	76.61	0.05	0.39	10.39	0.48
Mode 2	0.60	0.18	0.005	0.40	1.25	0.26

Table 4 Estimation of the parameters of the Beta and Gaussian mixtures for the Acidity data set

	Beta mixture			Gaussian mixture		
Parameters	P_j	s_j	m_j	P_j	μ_j	σ_j^2
Mode 1	0.58	123.56	0.54	0.42	34.56	0.78
Mode 2	0.59	4.33	0.13	0.41	6.24	0.27

and Gaussian mixtures, for the Enzyme and Acidity data sets, respectively. In both cases, it's clear that Beta and Gaussian mixtures fit the data. The final results of the estimations are given in tables 3 and 4. In order to rate the ability of Beta and Gaussian mixtures to fit the data, we used the Bayesian information criteria (BIC)[3] of Schwarz (1978). In fact, choosing a relevant model consists both of choosing its form f (Beta or Gaussian in our case) and the number of components M. By using a Bayesian approach, a way of selecting a model is to choose the one of highest posterior probability. Under regularity conditions, a classical way to approximate the posterior probability is to use BIC (Biernacki et al. 2000, Kass and Raftery 1995):

$$BIC = logp(\mathcal{X}|f, M, \hat{\Theta}_M) - \frac{N(M)}{2} log(N) \tag{26}$$

The first term is the familiar log-likelihood of the data given the model, computed at the value $\hat{\Theta}_M$ that maximizes this term. $N(M)$ is the number of parameters needed to specify a M-component mixture. Many simulation experiments have shown that the BIC approximation works well in practice

Table 5 BIC criteria values for several values of M when we consider the Beta mixture

M	1	2	3	4	5
Enzyme	−37.05	**−9.31**	−16.15	−26.09	−39.45
Acidity	−35.97	**−11.08**	−15.37	−25.19	−40.13

Table 6 BIC criteria values for several values of M when we consider the Gaussian mixture

M	1	2	3	4	5
Enzyme	−37.11	**−10.69**	−18.31	−26.22	−41.05
Acidity	−39.62	**−11.14**	−17.89	−25.97	−42.07

at practice (see for instance Roeder and Wasserman 1997, Biernacki et al. 2000). The values of the BIC criteria for both Beta and Gaussian mixture, for different values of M, are given in tables 5 and 6, respectively. According to these tables, the optimal number of components to fit the data, for both Gaussian and Beta mixtures, is $M = 2$. In this case, we can note that the BIC value when we use Beta mixture is higher than that for Gaussian mixture.

For the image processing application, we use the Bayesian approach to estimate the parameters of unimodal and multimodal SAR images histograms. In fact, many civil and military applications use SAR images, such as target detection, geological cartography, and oil spill and ship detection. The parameters estimated can then be used in segmentation, features extraction, pattern recognition, classification, etc. The use of the Beta mixture is motivated by the fact that the histograms of SAR images are always asymmetric. Besides, the data to be tested are defined in the compact support [0,255] which defines the different gray levels. The goal of this application is to compare the modeling capabilities of the Beta and Gaussian mixture, too. The comparison is based on BIC. Figures 11, 12 and 13 present some of these images with the real and the estimated histograms. The estimated parameters and the BIC values when we use both Beta and Gaussian mixtures are given in tables 7, 8 and 9.

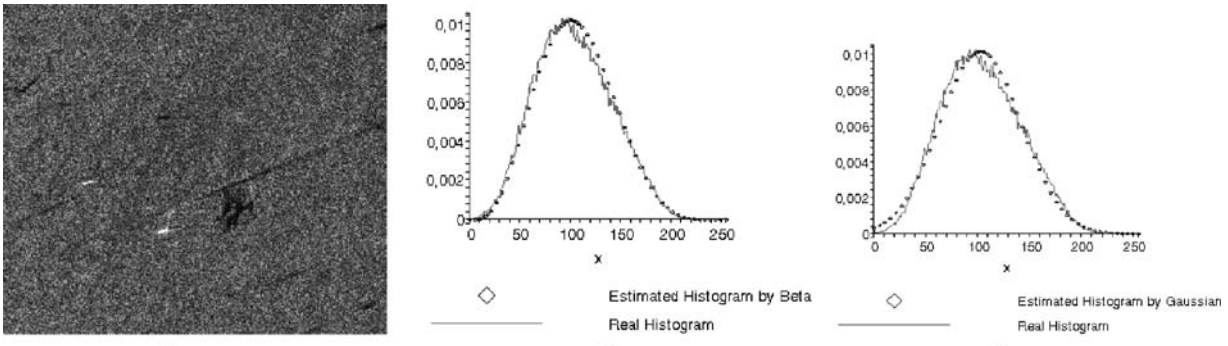

Fig. 11 (a) ERS SAR image of dimension 436 × 374. (b) Comparison between image histogram and Beta distribution. (b) Comparison between image histogram and Gaussian distribution

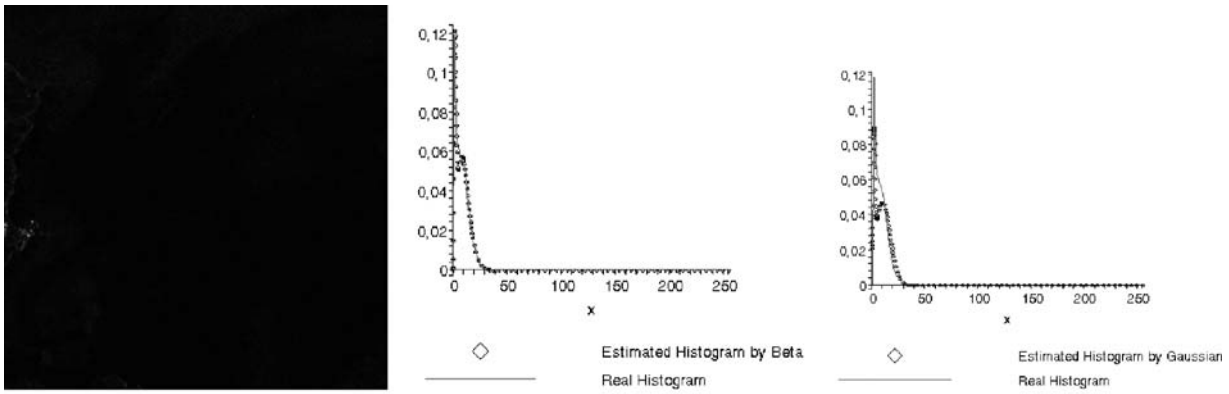

Fig. 12 (a) RADARSAT SAR image of dimension 1000 × 1000. (b) Comparison between image histogram and Beta mixture. (c) Comparison between image histogram and Gaussian mixture. © Canadian Space Agency 1997

4. Conclusion

This paper has established efficient Bayesian estimation of the parameters of a finite Beta mixture and provided a workable algorithm based on the Gibbs sampling. Our algorithm appear powerful to deal with the problems concerning the maximum likelihood approach. In fact, the estimation is based on simulation form the posterior distributions of Beta mixture model. These posteriors are given and justified through contextual and non-contextual evaluations. The non-contextual evaluation is based on artificial histograms. We have presented an example where the Bayesian estimation give good results while the ML method fails. The contextual evaluation model the class-conditional densities of well-known pattern recognition data sets and estimate the parameters of SAR images histograms. The difficulty of this model arises when the number of components M is unknown. The setting is, however, familiar, in that several solutions for this problem can be used. The two most prominent being Richardson and Green's (1997) reversible jump MCMC algorithm and Stephens (2000) birth-and-death process algorithm. Note that other purely numerical approaches can be used to estimate the parameters of the finite Beta mixture. However, we think that a desirable approach can combine both perspectives (Bayesian and numeric). Future work, will be devoted to the extension of this Bayesian approach to handle multivariate problems (finite Dirichlet mixture).

Acknowledgments The completion of this research was made possible thanks to the the Natural Sciences and Engineering Research Council of Canada, Heritage Canada and Bell Canada's support through its Bell University Laboratories R&D program. The authors would like to thank the anonymous referees for their helpful comments.

Notes

1. As an anonymous referee has noted, a Beta distribution would arise naturally in a Bayesian setting as the posterior distribution for the rate parameter of Bernoulli trials. A mixture of Beta distributions would naturally arise if each trial in the data set could be controlled by one of M rates,

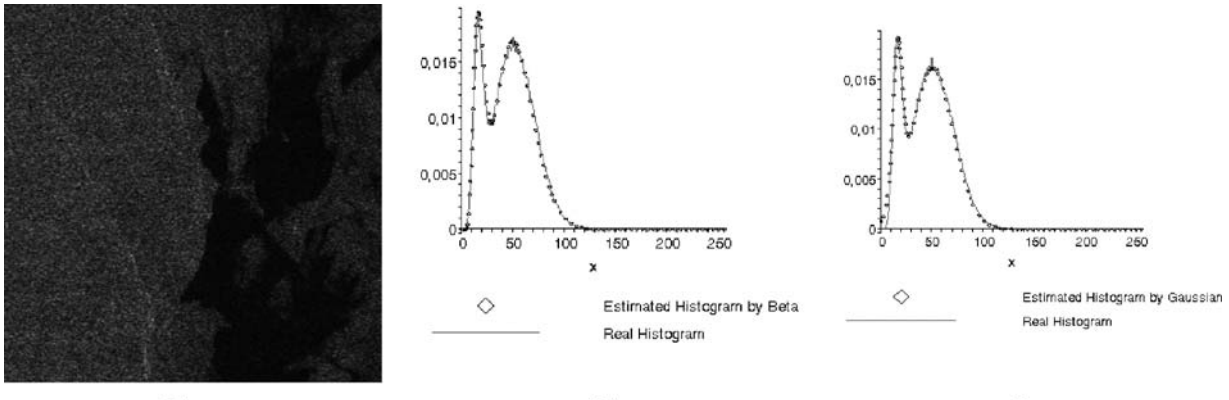

Fig. 13 (a) RADARSAT SAR image of dimension 512 × 512. (b) Comparison between image histogram and Beta mixture. (c) Comparison between image histogram and Gaussian mixture. ©Canadian Space Agency 1997

Table 7 Estimation of the parameters of the Beta and Gaussian mixtures for the histogram in Fig. 11

	Beta mixture			Gaussian mixture		
Parameters	P_j	s_j	m_j	P_j	μ_j	σ_j^2
Mode 1	1	10.67	0.59	1	102.42	39.54
BIC		−11.05			−11.83	

Table 8 Estimation of the parameters of the Beta and Gaussian mixtures for the histogram in Fig. 12

	Beta mixture			Gaussian mixture		
Parameters	P_j	s_j	m_j	P_j	μ_j	σ_j^2
Mode 1	0.27	565.33	0.01	0.13	2.27	0.86
Mode 2	0.73	86.41	0.04	0.87	6.91	7.35
BIC		−31.87			−32.54	

Table 9 Estimation of the parameters of the Beta and Gaussian mixtures for the histogram in Fig. 13

	Beta Mixture			Gaussian Mixture		
Parameters	P_j	s_j	m_j	P_j	μ_j	σ_j^2
Mode 1	0.22	139.95	0.07	0.17	16.27	4.59
Mode 2	0.78	30.91	0.21	0.83	50.53	20.47
BIC		−29.13			−29.79	

with the rate chosen according to the mixing proportions. Thus one interesting area in which our model could be useful would be in such Bernoulli trials where mixture of Beta are used as a prior. However, in this paper we consider the Beta as the parent distribution.

2. The use of a Metropolis-Hastings step rather than an accept-reject algorithm was suggested by an anonymous referee.
3. The use of the BIC criteria to rate the ability of Beta and Gaussian mixtures to fit the data was suggested by an anonymous referee.

References

McLachlan G. J. and Peel D. 2000. Finite Mixture Models. New York: Wiley.

Everitt B. S. and Hand D. J. 1981. Finite Mixture Distributions. Chapman and Hall, London, UK.

Roberts S. J. and Rezek L. 1998. Bayesian Approach to Gaussian Mixture Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11): 1133–1142.

Samuel K., Ng K. W., and Fang K. 1990. Symmetric Multivariate and Related Distributions. London/New York: Chapman and Hall.

Bouguila N., Ziou D., and Vaillancourt J. November 2004. Unsupervised Learning of a Finite Mixture Model Based on the Dirichlet Distribution and its Application. IEEE Transactions on Image Processing 13(11): 1533–1543.

Beckman R. J. and Tietjen G. L. 1978. Maximum Likelihood Estimation for the Beta Distribution. Journal of Statistics and Computational Simulation 7: 253–258.

Klieter G. 1992. Bayesian Diagnosis in Expert Systems. In AIJ92.

Lee J. C. and Lio Y. L. 1999. A Note on Bayesian Estimation and Prediction for the Beta-binomial Model. Journal of Statistical Computation and Simulation (63): 73–91.

Dempster A. P., Laird N. M., and Rubin D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, B 39: 1–38.

McLachlan G. J. and Krishnan T. 1997. The EM Algorithm and Extensions. New York: Wiley.

Robert C. P. and Casella G. 1999. Monte Carlo Statistical Methods. Springer-Verlag.

Diebolt J. and Robert C. P. 1994. Estimation of Finite Mixture Distributions Through Bayesian Sampling. Journal of the Royal Statistical Society, B 56(2): 363–375.

Tsung I. L., Jack C. L., and Huey F. N. 2004. Bayesian Analysis of Mixture Modeling Using the Multivariate t Distribution. Statistics and Computing 14: 119–130.

Tsionas E. G. 2004. Bayesian Inference for Multivariate Gamma Distributions. Statistics and Computing 14: 223–233.

Celeux G. and Diebolt J. 1985. The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. Computational Statistics Quarterly 2(1): 73–82.

Celeux G. and Diebolt J. 1992. A Stochastic Approximation Type EM Algorithm for the Mixture Problem. Stochastics and Stochastics Reports 41: 119–134.

Escobar M. and West M. 1995. Bayesian Prediction and Density Estimation. Journal of the American Statistical Association 90: 577–588.

Marin J. M., Mengersen K., and Robert C. P. 2004. Bayesian modeling and inference on mixtures of distributions. In D. Dey and C.R. Rao, editors, Handbook of Statistics 25. Elsevier-Sciences.

Robert C. P. and Rousseau J. 2002. A Mixture Approach To Bayesian Goodness of Fit. Technical Report 02009, Cahier du CEREMADE, Université Paris Dauphine.

Casella G., Mengersen K., Robert C., and Titterington D. 2000. Perfect Slice Samplers for Mixtures of Distributions. Journal of the Royal Statistical Society, B 64(4): 777–790.

Bezdek J. C. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Fielitz B. D and Myers B. L 1975. Estimation of Parameters in the Beta Distribution. Decision Sciences 6: 1–13, 1975.

Crawford S. L. 1994. An application of the Laplace Method to Finite Mixture Distributions. Journal of the American Statistical Association 89: 259–267.

Schwarz G. 1978. Estimating the Dimension of a Model. Annals of Statistics 6: 461–464.

Biernacki C., Celeux G. and Govaert G. 2000. Assessing a Mixture Model for Clustering with the Integrated Complete Likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(7): 719–725.

Kass R. E. and Raftery A. E. 1995. Bayes Factor. Journal of the American Statistical Association 90: 733–795.

Roeder K. and Wasserman L. 1997. Practical Bayesian Density Estimation Using Mixture of Normals. Journal of the American Statistical Association 92: 894–902.

Richardson S. and Green P. J. 1997. On Bayesian Analysis of Mixtures with an Unknown Number of Components (With Discussion). Journal of the Royal Statistical Society, B 59: 731–792.

Stephens M. 2000. Bayesian Analysis of mixture Models with an Unknown Number of Components: An Alternative to reversible Jump Methods. Annals of Statistics 28:40–74.