Rapporto n° 201

Modeling distributions on a bounded support via finite mixtures of mode-parameterized beta and gamma densities

Luca Bagnato, Antonio Punzo

Dicembre 2010

Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali
Università degli Studi di Milano Bicocca
Via Bicocca degli Arcimboldi 8 - 20126 Milano - Italia
Tel +39/02/64483102/3 - Fax +39/2/64483105
Segreteria di redazione: Andrea Bertolini

Modeling distributions on a bounded support via finite mixtures of mode-parameterized beta and gamma densities

Luca Bagnato · Antonio Punzo

Abstract This paper address the problem of estimating a density, defined on a bounded interval, exploiting a general and natural form of finite mixture of distributions. To this end, subclasses of unimodal beta and gamma densities are used as components in the mixture. These belong to the Pearson family of curves, whose definition consent mode-parameterized densities. The mode is the natural parameter since mixtures of distributions are strictly related to the concept of multimodality. The EM algorithm for maximum likelihood estimation of the mixture parameters, is also described. For this algorithm, the choice of good starting values plays an important role. Here we propose a simple and *ad hoc* initialization strategy, based on bump-hunting; its performance, in comparison with random initialization, is also evaluated by some simulation experiments. Finally, two real data sets are considered, in order to appreciate the advantages of the adopted parameterization for both components.

Keywords Finite mixtures of densities · Pearson system · EM algorithm · Initialization strategies · Bump hunting · Recovery rates · Number of first births

Luca Bagnato
Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali, Università di Milano-Bicocca (Italy)
Tel.: +39-02-64483186
Fax: +39-02-64483105
E-mail: luca.bagnato@unimib.it

Antonio Punzo
Dipartimento Impresa, Culture e Società, Università di Catania (Italy)
Tel.: +39-095-7537732
Fax: +39-095-7537610
E-mail: antonio.punzo@unict.it

1 Introduction

This paper considers estimation of a probability density function, with bounded support S, via finite mixtures of univariate densities. Some comprehensive and exhaustive reviews of this class of models are available in Everitt and Hand (1981), Redner and Walker (1984), Titterington et al (1985), Lindsay (1995) and McLachlan and Peel (2000). Roughly speaking, according to Titterington et al (1985, p. 2), there are two broad motivations for using finite mixture of univariate densities. Firstly, from a *direct* viewpoint, we may believe in the existence of g underlying groups such that each of the n observations x_i is made to belong to one of these; logically, we do not observe the source of x_i. Secondly, from an *indirect* viewpoint, a finite mixture of densities could simply be used as a mathematical device in order to provide a flexible and tractable form of analysis; in these terms, it represents a sort of semiparametric compromise between a single ($g = 1$) parametric density and a nonparametric kernel method of density estimation, represented in the case $g = n$ (see Jordan and Xu 1995, McLachlan and Peel 2000, p. 8, Titterington et al 1985, p. 29).

Most of the work published is concerned with mixture of normal densities. Unfortunately, while using Gaussian components in the mixture is appropriate for fitting densities with unbounded supports, it is not adequate for densities with compact or bounded from one end only supports as it causes *boundary bias*; that is, allocation of probability mass outside the theoretical support S. A simple remedy is to use mixture components defined on S (see Chen, 1999, 2000, for an analogous of this suggestion in the kernel density estimation context). Motivated by this consideration and according to Chen (1999, 2000), we suggest using gamma components for bounded from one end only supports, and beta components for compact ones.

In analogy with the normal case, we have chosen to only focus attention on unimodal beta and gamma densities, adopting a mode-based parameterization. This parameterization arises naturally recalling that such distributions, in addition to the normal one, belong to the *Pearson system* of density curves (Pearson, 1902a,b). A general framework for finite mixture of unimodal distributions is therefore provided which includes normal-mixtures when $S = \mathbb{R}$. The choice of using "mode-parameterized" components is justified, but above all natural, if one thinks that the most striking feature of a mixture density curve is often that of multimodality. Indeed, as highlighted in Titterington et al (1985) and McLachlan and Basford (1988), many papers in applied fields talk not in terms of mixtures but of multimodal distributions; examples are the articles of Murphy (1964) and Brazier et al (1983) referring to bimodality rather than to mixtures.

As usual, in order to estimate the parameters of the mixture by maximum likelihood (ML), the EM algorithm (Dempster et al, 1977) is taken into account. Despite its good properties, as well-known, it needs to a good initialization method in order to be sure to find the global maximum of the likelihood.

As far as we know to date, different initialization strategies have only been investigated for mixtures of multivariate normal distributions, while random initialization is commonly used otherwise. Unfortunately, this procedure could induce the EM algorithm to converge to different local maxima any time the algorithm is executed. Motivated by this consideration, an *ad-hoc* initialization procedure, based on bump-hunting (BH) and coherent with the adopted parameterization, is proposed and evaluated by some numerical experiments.

The paper can be schematically summarized as follows: in Section 2 the general framework for finite mixtures of unimodal distributions defined on S is introduced; details of the parameterization used for the gamma and beta components are also provided in Section 2.1 and Section 2.2, respectively. The EM-algorithm for ML estimation of the mixture-parameters is discussed in Section 3, with further details given in Appendix A, while the proposed BH-initialization strategy is described in Section 4. In Section 5 some simulation experiments are illustrated in order to evaluate the behavior of the initialization proposal in various situations arising in practice. Two real applications are also described in Section 6 to appreciate the advantages of the adopted parameterization for both the components, beta (see Section 6.1) and gamma (see Section 6.2). Finally, in Section 7, conclusions are drawn.

2 A general framework for finite mixture of distributions

In requiring that the component densities should all belong to the same parametric family, a general finite mixture density function f could assume the form

$$f\left(x;\boldsymbol{\pi},\boldsymbol{m},\boldsymbol{v}\right)=\sum_{j=1}^{g}\pi_j f_j\left(x;m_j,v_j\right),\quad x\in S, \tag{1}$$

where

- f_j is the unimodal *component* density defined on S and belonging to some convenient parametric family;
- $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_g)$ is the g-dimensional vector of the mixture *weights* $\pi_j\in(0,1)$, with $\sum_{j=1}^{g}\pi_j=1$;
- $\boldsymbol{m}=(m_1,\ldots,m_g)$ is the g-dimensional vector of the *modes* $m_j\in S$ of f_j;
- $\boldsymbol{v}=(v_1,\ldots,v_g)$ is a g-dimensional vector containing parameters $v_j>0$ governing the concentration of f_j around the mode m_j.

Thus, there are $3g-1$ unconstrained parameters to be estimated. Of course, as also underlined by Izenman (2008, p. 103), there is no guarantee that a mixture of unimodal densities will produce a multimodal density with the same number of modes as there are densities in the mixture; similarly, there is no guarantee that those individual modes m_j will remain at the same locations in (1). Indeed, the shape of the mixture distribution depends upon both the spacings of the modes and the relative shapes of the component distributions. Nevertheless, we retain that for well-separated components, the values of m_j will well-approximate the location of the mixture-modes in (1).

More specifically, motivated by the works of Chen (1999, 2000), we have chosen to adopt mode-parameterized unimodal beta densities for compact supports and unimodal gamma densities for bounded from one end only ones. The adjective "unimodal" is useful to highlight the subclass of beta and gamma densities on which attention is focused. However, other distributions may be used if they can be parameterized according to m_j. In doing so, it is natural to take into account the *Pearson system* of density curves (Pearson, 1902a,b). Indeed, these curves originate from a differential equation having m_j between its parameters; examples are the beta density for a compact support; the gamma and lognormal densities for a bounded from one end only support; the normal, Student t and Cauchy densities for an unbounded support (see Elderton and Johnson 1969, Johnson and Kotz 1970, and Kendall and Stuart 1958, for further details). Given the importance of normals in the mixture context, it is worthwhile noting that if S is unbounded, we could use normal components in (1), with $m_j = \mu_j$ and $v_j = \sigma_j$ (or, equivalently, $v_j = \sigma_j^2$), obtaining the popular finite mixture of normal distributions.

In the following, some properties of the adopted distributions are sketched.

2.1 Bounded from one end only support: $S = [a, \infty)$

The following class of gamma components

$$f_j\left(x; m_j, v_j\right) = \frac{\left(x-a\right)^{\frac{m_j-a}{v_j}} e^{-\frac{x-a}{v_j}}}{v_j^{\frac{m_j-a}{v_j}+1} \Gamma\left(\frac{m_j-a}{v_j}+1\right)}, \qquad a \leq x < \infty, \tag{2}$$

can be considered in (1) when $S = [a, \infty)$, with m_j and v_j satisfy the above-mentioned conditions. The expectation of (2) is easily obtained as:

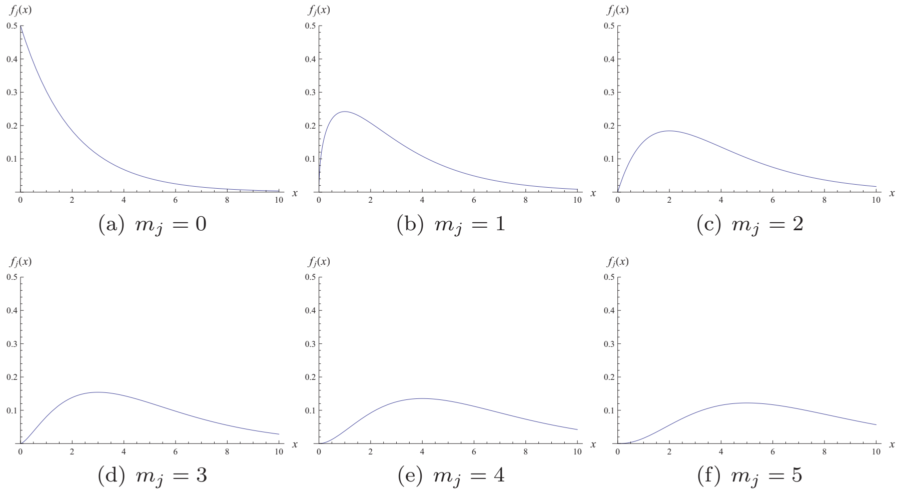
$$E\left(X\right) = m_j + v_j .$$

An eminent feature of gamma components is that their shape changes according to the value of m_j; this is shown by a set of gamma densities displayed in Fig. 1. The variance of a random variable ($r.v.$) with density function (2), from the standard theory on gamma distribution, is

$$v_j^2 + \left(m_j - a\right) v_j . \tag{3}$$

The last expression, analyzed as a function of m_j, is a straight line with a positive slope v_j; consequently, fixed v_j, the variability increases in line with the value of m_j. Conversely, fixing m_j in (3), the variance increases if v_j increases, confirming that v_j governs the spread of the distribution. The effect of varying v_j, fixed the mode m_j, is illustrated in Fig. 2.

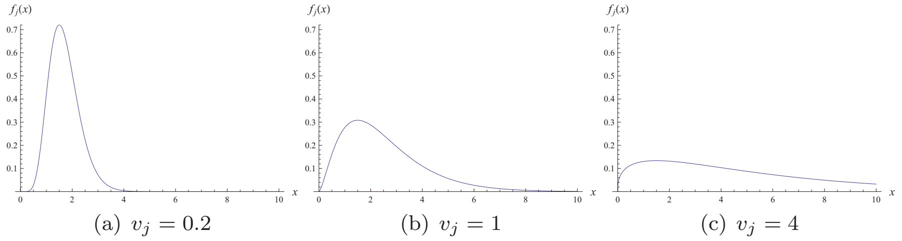
Finally, note that $(-\infty, b]$ is also part of the bounded from one end only supports. If the underlying density f is defined in $S = (-\infty, b]$, then it is



(a) $m_j = 0$ (b) $m_j = 1$ (c) $m_j = 2$

(d) $m_j = 3$ (e) $m_j = 4$ (f) $m_j = 5$

Fig. 1: Gamma components defined in $S = [0, \infty)$, with $v_j = 2$



(a) $v_j = 0.2$ (b) $v_j = 1$ (c) $v_j = 4$

Fig. 2: Gamma components defined in $S = [0, \infty)$, with $m_j = 1.5$

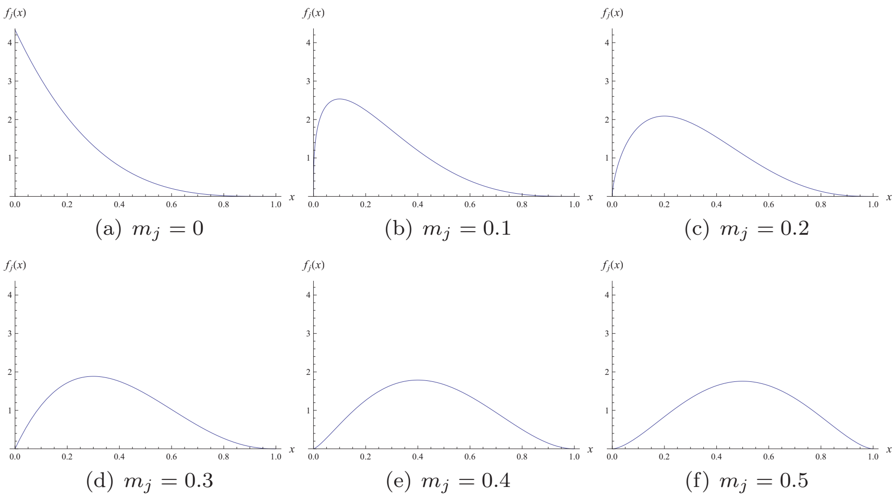
sufficient to consider the following gamma components

$$f_j\left(x; m_j, v_j\right) = \frac{(b-x)^{\frac{b-m}{v_j}}\, e^{-\frac{b-x}{h}}}{v_j^{\frac{b-m}{v_j}+1}\Gamma\left(\frac{b-m}{v_j}+1\right)}, \qquad -\infty < x \leq b. \tag{4}$$

2.2 Compact support: $S = [a, b]$

When $S = [a, b]$, in (1) it is possible to consider the class of beta components

$$f_j\left(x; m_j, v_j\right) = \frac{(x-a)^{\frac{m_j-a}{v_j(b-a)}}\,(b-x)^{\frac{b-m_j}{v_j(b-a)}}}{(b-a)^{\frac{v_j+1}{v_j}}\ \mathrm{B}\left[\frac{m_j-a}{v_j\,(b-a)}+1, \frac{b-m_j}{v_j\,(b-a)}+1\right]}, \qquad a \leq x \leq b, \tag{5}$$



(a) $m_j = 0$ (b) $m_j = 0.1$ (c) $m_j = 0.2$

(d) $m_j = 0.3$ (e) $m_j = 0.4$ (f) $m_j = 0.5$

Fig. 3: Beta components defined in $S = [0, 1]$, with $v_j = 0.3$

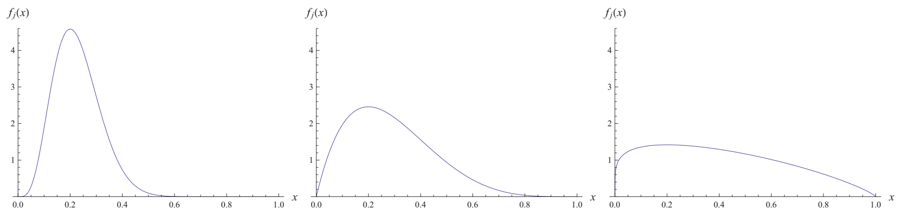
with, as before, $m_j \in S$ and $v_j > 0$. The same parameterization in conveniently re-adapted in the discrete case by Punzo (2010) and Mazza and Punzo (2011) to define two different nonparametric estimators. The expectation of (5) is:

$$E(X) = \frac{m_j + (a+b)\, v_j}{2v_j + 1}. \tag{6}$$

In analogy with the gamma components (2)-(4), the beta density shape changes according to the value of m_j; this is shown by a set of beta components displayed in Fig. 3. Note that the variance of a *r.v.* with density function (5), from the standard theory on beta distribution, is

$$\frac{v_j \left[(m_j - a) + v_j (b-a)\right]\left[(b - m_j) + v_j (b-a)\right]}{(2v_j + 1)^2 (3v_j + 1)}. \tag{7}$$

The expression (7) is a parabola, if it is analyzed as a function of m_j, with maximum in correspondence to $m_j = (a+b)/2$ (central point of S). In other words, fixed v_j, the variability decreases as $|m_j - (a+b)/2|$ increases. On the other hand, fixing m_j in (7), the variance increases if v_j increases, confirming the previously-mentioned requirement. The effect of varying v_j, fixed m_j, is illustrated in Fig. 4. In more detail, the limit of (7), as v_j tends to zero, is zero, while as v_j becomes large, the limit is $(b-a)^2/12$, that is the variance of a uniform distribution defined on $[a, b]$ (note that the beta distribution (5) converges to a uniform distribution when $v_j \to \infty$).



(a) $v_j = 0.05$ (b) $v_j = 0.2$ (c) $v_j = 1$

Fig. 4: Beta components defined on $S = [0, 1]$, with $m_j = 0.2$

3 Estimation by the EM algorithm

In the conceptualization of a mixture model, hypothesizing knowing the number of groups g in advance, the observed data $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_n)$ are usually augmented by introducing, for each observation x_i, the g-dimensional latent component-indicator vector $\boldsymbol{z}_i$ in which the single element is defined as follows

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ comes from group } j \\ 0 & \text{otherwise.} \end{cases}$$

The complete-data vector is therefore declared to be $\boldsymbol{x}_c = (\boldsymbol{x}', \boldsymbol{z}')'$, where $\boldsymbol{z} = (\boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_n)$. Here it is assumed that each vector z_i is independent and that all the observations x_i have been completely recorded. So, the likelihood function for the complete-data is given by

$$L_c(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{v}) = \prod_{i=1}^{n} \prod_{j=1}^{g} \left[\pi_j f_j(x_i; m_j, v_j)\right]^{z_{ij}} .$$

Consequently, the complete-data log-likelihood is given by

$$l_c(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{v}) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \left[\ln \pi_j + \ln f_j(x_i; m_j, v_j)\right] . \tag{8}$$

Equation (8) can be iteratively maximized by the EM algorithm (Dempster et al, 1977). It proceeds in the following two steps:

E-step: On the $(k+1)$th iteration, $k = 0, 1, \ldots$, the E-step simply requires the calculation of the conditional expectation of Z_{ij} − the random variable corresponding to z_{ij} − given the observed data $\boldsymbol{x}$ and the parameter estimates from the M-step arising from the kth iteration

$$\begin{aligned} z_{ij}^{(k+1)} &= E\left(Z_{ij} \,\middle|\, x_i, \boldsymbol{\pi}^{(k)}, \boldsymbol{m}^{(k)}, \boldsymbol{v}^{(k)}\right) \\ &= P\left(Z_{ij} = 1 \,\middle|\, x_i, \boldsymbol{\pi}^{(k)}, \boldsymbol{m}^{(k)}, \boldsymbol{v}^{(k)}\right) \\ &= \frac{\pi_j^{(k)} f_j\left(x_i; m_j^{(k)}, v_j^{(k)}\right)}{\sum_{j=1}^{g} \pi_j^{(k)} f_j\left(x_i; m_j^{(k)}, v_j^{(k)}\right)} \end{aligned} \tag{9}$$

$i = 1, \dots, n$ and $j = 1, \dots, g$. The quantity $z_{ij}^{(k+1)}$ in (9) is the posterior probability that x_i belongs to the jth component of the mixture.

M-step: On the $(k+1)$th iteration, the M-step requires the global maximization of (8) with respect to $\boldsymbol{\pi}$, $\boldsymbol{m}$, and $\boldsymbol{v}$, replacing z_{ij} by $z_{ij}^{(k+1)}$; the aim is to obtain the updated estimates $\boldsymbol{\pi}^{(k+1)}$, $\boldsymbol{m}^{(k+1)}$ and $\boldsymbol{v}^{(k+1)}$. For the finite mixture models, the updated estimates $\pi_j^{(k+1)}$ are calculated independently of the updated estimates $m_j^{(k+1)}$ and $v_j^{(k+1)}$. Specifically, if the z_{ij} were observable, then the complete-data ML-estimate of π_j would be given simply by

$$\widehat{\pi}_j = \frac{\sum_{i=1}^{n} z_{ij}}{n}, \quad j = 1, \dots, g. \tag{10}$$

As the E-step simply involves replacing each z_{ij} with its current conditional expectation $z_{ij}^{(k+1)}$ in the complete-data log-likelihood, the updated estimate of π_j is given by replacing each z_{ij} in (10) by $z_{ij}^{(k+1)}$ to give

$$\pi_j^{(k+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(k+1)}}{n}, \quad j = 1, \dots, g.$$

Concerning the updating of $\boldsymbol{m}$ and $\boldsymbol{v}$ on the $(k+1)$th iteration for the M-step, it can be seen from (8) that $\boldsymbol{m}^{(k+1)}$ and $\boldsymbol{v}^{(k+1)}$ are obtained as an appropriate root of

$$\begin{cases} \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(k+1)} \frac{\partial \ln f_j\left(x_i; m_j, v_j\right)}{\partial \boldsymbol{m}} = \mathbf{0} \\ \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(k+1)} \frac{\partial \ln f_j\left(x_i; m_j, v_j\right)}{\partial \boldsymbol{v}} = \mathbf{0} \end{cases} \tag{11}$$

The solution of (11) exists in closed form when the components are normals; otherwise, maximization can only be carried out numerically. Details on the derivatives in (11), for the components (2) and (5), are given in Appendix A.

The final parameter estimates $\widehat{\pi}_j$, $\widehat{m}_j$ and $\widehat{v}_j$ are so obtained, starting from a set of initial values $\pi_j^{(0)}$, $m_j^{(0)}$ and $v_j^{(0)}$, $j = 1, \dots, g$, alternatively repeating the E- and M-steps until the difference in two consecutive values of the log-likelihood in (8) is retained negligible (for further details on this algorithm, applied to finite mixture models, see McLachlan and Peel, 2000; McLachlan and Krishnan, 2007).

It is worth noting, that while the convergence of the algorithm to a local maximum of the log-likelihood is guaranteed, the identification of the global maximum cannot be assured. Indeed, the solution found by the EM algorithm strongly depends on the choice of the initial parameters. In order to address this issue, we propose an initialization strategy based on Bump-Hunting (BH) which will be described in Section 4.

4 Using Bump-Hunting to initialize the EM algorithm

As mentioned in Section 3, the EM algorithm is sensitive to the problem of choosing the initial values of the parameters to be estimated. If these initial values are inappropriately selected, it may lead to an unsatisfactory estimation of the mixture. In an attempt to resolve this problem, several methods are reported in literature although, so far, they have only been investigated for data from mixtures of multivariate normal distributions (see Biernacki et al, 2003, for an overview of simple initialization strategies). The random initialization that consists in initializing the EM algorithm from a random position, is probably the one most used. An extension of this simple strategy consists in repeating it t times from different random positions and selecting the solution maximizing the likelihood among those t runs. The resulting algorithm will be denoted hereafter as "tR-EM". Unfortunately, final estimates from both methods could be different every time the algorithm is executed. To avoid these occurring, some authors use the result of the k-means algorithm – that in our context should be more correctly named as "g-means algorithm" – to set the initial cluster centers $\overline{x}_j$ and the initial cluster variances s_j^2, $j = 1, \ldots, g$. Nevertheless, the advantage of using this strategy is minimal, since this does not assure that the k-means algorithm will not itself be trapped in local minimum decisions (Dempster et al, 1977; Khan and Ahmad, 2004). Furthermore, it can not be adopted with our parameterization, since for beta components, it is not possible to obtain the values of $m_j^{(0)}$ and $v_j^{(0)}$ starting from the values of $\overline{x}_j$ and s_j^2, $j = 1, \ldots, g$; equations (6) and (7) in fact, once put equal respectively to $\overline{x}_j$ and s_j^2, can not be simultaneously inverted in order to determine m_j and v_j.

It is our belief, also supported by Meilă and Heckerman (2001), that we should not expect to find an "optimal" initialization strategy that outperforms all the others on all data sets. Thus, we simply propose an initialization method that we retain *ad hoc* for the model and the parameterization used, and that we hope works well for large classes of situations arising in practice (the answer to this question will be provided in Section 5 by numerical experiments for various data sets).

4.1 Some preliminary considerations

The underlying idea for our proposal originates from a common belief in the mixture framework. Indeed, often, as underlined by Titterington et al (1985, pp. 49–50), unless the separation between components is enough to manifest multimodality, there is not sufficient evidence in the data to confidently reject the pure component hypothesis. On the contrary, we also believe that a bump (a part lying between two points of inflection in a probability density curve without straight parts, and that is concave when viewed from below) even if within it there is not a local maximum (a mode), indicates some feature of the random variable requiring an explanation. In confirmation of our conviction,

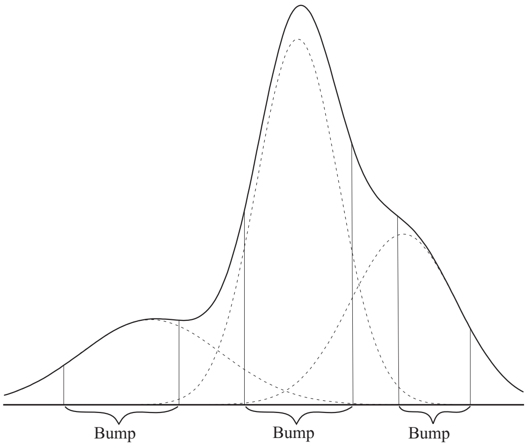


Fig. 5: Bumps associated with a mixture of three gaussian densities

it is sufficient to note that the existence of more than one bump has been discussed by Cox (1966) as a "descriptive feature likely to indicate mixing of components". To agree with our remarks, from a direct point of view, it is sufficient to recognize that bumps arise when unimodal distributions are mixed and each bump can be interpreted as a group. Figure (5) displays how effectively bumps can be linked to mixture components. This way of thinking also has an advantage from an indirect point of view where the interest is often focused on the flexibility of the density estimator and consequently, on the goodness-of-fit with the empirical density. For example, it is easy to agree with the observation that, if a unimodal density presents two bumps, it will be difficult for it to be well-fitted by a classic unimodal density (note that standard parametric densities are characterized by a single bump) should one not consider the mix with a further unimodal component.

In order to detect bumps, our approach readjusts the idea of "critical smoothing" used by Silverman (1981) to investigate multimodality. Consider the kernel density estimator

$$\hat{f}\left(x,h\right)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right), \tag{12}$$

where K is a kernel function, which we shall assume throughout to be the normal density function, and h is the bandwidth controlling the amount of smoothing. Suppose we fix the number of components g and that we want to find a density estimate (12) that highlights g bumps and consequently g components. Using and extending the results of Silverman (1981), we can define the g-critical bandwidth h_g by

$$h_g=\inf\left\{h:\hat{f}\left(\cdot,h\right)\ \text{has at most } g \text{ bumps}\right\}. \tag{13}$$

Thus, h_g identifies the most roughly kernel density estimate which highlights g bumps. Since $v(h)$, the number of sign changes in $\hat{f}^{(2)}(\cdot,h)$, is a right continuous decreasing function of h (Silverman, 1981), then also $b(h)$, the number of bumps, is a right continuous decreasing function of h. From a practical viewpoint, h_g can be obtained through a simple binary search procedure. In particular, we have that an interval (h_1,h_2), in which h_g is known to fall, can be halved in length by checking whether the value $(h_1+h_2)/2$ leads to an estimate $\hat{f}$ with more then g bumps.

4.2 Starting values for the EM algorithm

Naturally, once the value of g in (1) is fixed, the BH binary search procedure can be used to find h_g in (13). Substituting this value in (12), the starting values $m_1^{(0)},\dots,m_g^{(0)}$ can be obtained in a natural way by considering the maximum values of $\hat{f}(x,h_g)$ in each bump. It is important to note, that in the internal part of the jth bump B_j, we have $\hat{f}^{(2)}(x,h_g)<0$. Only at the boundaries of B_j we have $\hat{f}^{(2)}(x,h_g)=0$. Then, the generic starting value $m_j^{(0)}$ will be obtained as

$$m_j^{(0)}=\underset{x\in B_j}{\operatorname{argmax}}\left\{\hat{f}\left(x,h_g\right)\right\},\quad j=1,\dots,g. \tag{14}$$

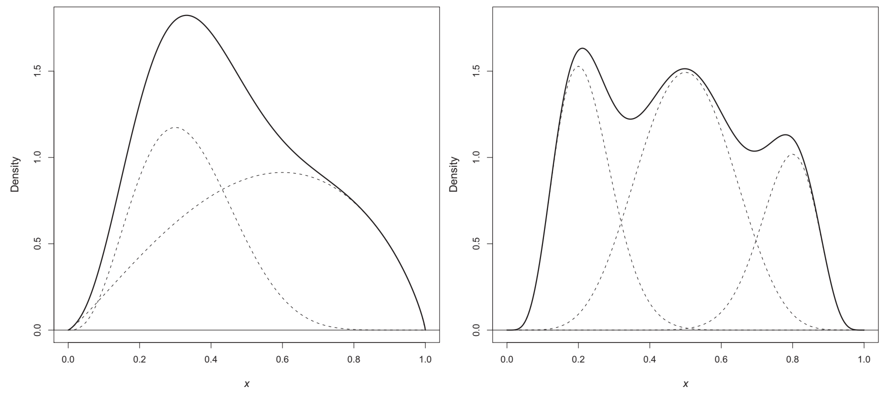
As regards the other two sets of parameters, they can be easily obtained by taking advantage of the set of initial modes in (14). In analogy with the k-means procedure, once S is fixed, the sample observations $x_1,\dots,x_n$ can be assigned to the g groups according to their closeness with the modes in (14). In detail, the support S can be partitioned in g disjoint subintervals S_j delimited by the $g-1$ internal cutoff points $\left(m_l^{(0)}+m_{l+1}^{(0)}\right)/2$, $l=1,\dots,g-1$. Naturally, $m_j^{(0)}$ will belong to the jth interval S_j. Thus, it is simple to define $z_{ij}^{(0)}=1$ if $x_i\in S_j$, and $z_{ij}^{(0)}=0$ otherwise. Note, that in these terms, the procedure described so far can be considered as an alternative method to the k-means for non-hierarchical clustering; it can also be easily extended in the multivariate context. The initial values for the weights can be so obtained as

$$\pi_j^{(0)}=\frac{\sum_{i=1}^{n}z_{ij}^{(0)}}{n},\quad j=1,\dots,g.$$

Finally, since we know the functional form f_j of each component (beta or gamma), the starting values for v_j, $j=1,\dots,g$, can be obtained, by ML, as

$$v_j^{(0)}=\underset{v_j>0}{\operatorname{argmax}}\left\{\prod_{i=1}^{n}\left[f_j\left(x_i;m_j^{(0)},v_j\right)\right]^{z_{ij}^{(0)}}\right\}.$$

The EM algorithm initialized according to the procedure described above will be hereafter denoted as BH-EM.



(a) Data B1

(b) Data B2

Fig. 6: Finite mixtures of standard beta densities used to generate data

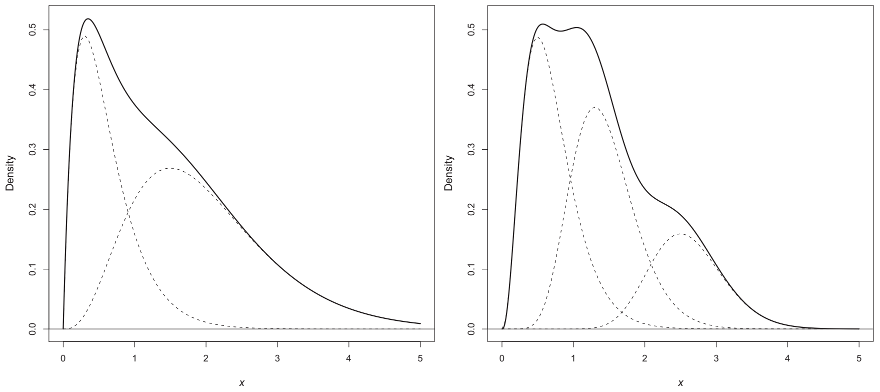
5 Experimental results

Once the initialization procedure has been described, it is important to evaluate its performance in various situations arising in practice. To address this issue, we use two convenient artificial data sets for each type of component in the mixture, and compare the behavior of the BH-EM, with respect to the tR-EM with $t = 1, \dots, 10$, in directing the EM algorithm towards the correct estimates.

To perform these numerical experiments, we use the `R` environment. The `R` functions necessary to implement graphics and estimates are available from the authors upon request. Regarding the stopping rule, we have chosen to stop the EM algorithm with the number of iterations. We do not use stopping criteria based on the relative change of the estimates or log-likelihood because the slow convergence of the EM makes such criteria hazardous (see Lindsay 1995 and McLachlan and Peel 2000 for more on stopping criteria). More specifically, we consider 1000 iterations that, according to Biernacki et al (2003), represent a good compromise.

As regards the beta components, two types of data in $[0, 1]$ have been generated from the densities displayed in Fig. 6. Data B1 arise from a unimodal two-components beta mixture, with a further bump, characterized by parameters $\pi_1 = 0.4$, $\pi_2 = 0.6$, $m_1 = 0.3$, $m_2 = 0.6$, $v_1 = 0.1$ and $v_2 = 0.5$ (see Fig. 6(a)), while data B2 arise from a trimodal three-components beta mixture with parameters $\pi_1 = 0.3$, $\pi_2 = 0.5$, $\pi_3 = 0.2$, $m_1 = 0.2$, $m_2 = 0.5$, $m_3 = 0.8$, $v_1 = 0.04$, $v_2 = 0.08$ and $v_3 = 0.04$ (see Fig. 6(b)).

Likewise, for the gamma mixtures, two types of data in $[0, \infty)$ have been considered. The distributions used to generate them are graphically represented in Fig. 7. Data G1 arise from a unimodal two-components gamma mixture, with a further bump, characterized by parameters $\pi_1 = 0.4$, $\pi_2 = 0.6$, $m_1 = 0.3$, $m_2 = 1.5$, $v_1 = 0.3$ and $v_2 = 0.5$ (see Fig. 7(a)), while data G2 arise



(a) Data G1

(b) Data G2

Fig. 7: Finite mixtures of gamma densities in $[0, \infty)$ used to generate data

Table 1: Percentage of times that each initialization method induces the EM algorithm towards correct estimates

Data	1R-EM	2R-EM	3R-EM	4R-EM	5R-EM	6R-EM	7R-EM	8R-EM	9R-EM	10R-EM	BH-EM
B1	93	98	100	100	100	100	100	100	100	100	100
B2	34	48	55	64	68	72	81	83	86	86	88
G1	62	83	91	92	94	98	98	98	99	99	99
G2	77	86	90	93	95	95	96	96	96	96	96

from a bimodal three-components gamma mixture, with a further bump, having parameters $\pi_1 = 0.4$, $\pi_2 = 0.4$, $\pi_3 = 0.2$, $m_1 = 0.5$, $m_2 = 1.3$, $m_3 = 2.5$, $v_1 = 0.2$, $v_2 = 0.14$ and $v_3 = 0.1$ (see Fig. 7(b)).

For each type of data, we have generated 100 samples of size $n = 300$. In our experiments, the algorithms have been run with two components for B1 and G1 and with three components for B2 and G2. Table 1 for each data structure, provides the percentage of time that the algorithm converges towards the best estimates. These "best" estimates are obtained initializing the EM-algorithm with the true but unknown values of the parameters. From Table 1 it appears that BH-EM never performs worse than the 10R-EM and, for data B2, it behaves even better; moreover, computational times of the BH-EM are, on average, similar to the 3R-EM but with a better performance. Further experiments, whose results are not reported here, have highlighted that the performance of the BH-EM does not vary at the increase of g. The same reasoning does not hold for the tR-EM whose performance decreases when g increases. This is probably due to the fact that random initialization tends to produce initial values $m_j^{(0)}$, $j = 1, \dots, g$, that are uniformly distributed on S even if the true but unknown values are not so.

6 Real applications

In this section, in order to appreciate the advantages of both the model and the adopted parameterization, the mode-parameterized mixtures will be applied to two data sets taken from two different real fields. In reality, a large number of applications of beta and gamma mixtures, although they are not parameterized as we suggest, can already be noted in literature; for example, Ji et al (2005) study the correlations of gene-expression levels in bioinformatics by finite mixtures of beta densities, while Mayrose et al (2005) use finite mixtures of gamma distributions to better describe the among-site rate variation, characteristic of molecular sequence evolution. In particular, in Ji et al (2005) the importance of knowing the position of the modes so as to make inference about the behavior of genes belonging to the corresponding component, is board to our attention. For this reason we looked for fields where our parameterization could be useful. Moreover, in choosing the applications, we have tried to find fields of application where, in the authors' knowledge, beta and gamma mixtures have not yet been adopted.

6.1 Recovery rates on Italian bank loans

The credit risk analysis is the first of the fields of application taken into account. More specifically, the example illustrated here, focuses on recovery risk of bank loans. The importance in analyzing such a risk is justified by the Basel II Accord which requires its measurement (Basel Comittee on Banking Supervision, 2004, paragraph 286). In particular, it is relevant to estimate the distribution of the random variable $X=$"recovery rate". Among the many approaches existing in related literature for the calculation of recovery rates (see Altman et al, 2005, for a survey), the proposal by Calabrese and Zenga (2008) is considered here; it is based on the concept of "total exposure" and constrains the rates to assume values in the compact interval $[0,1]$.

The Bank of Italy's data on 149378 bank loan recovery rates are taken into account here (see Banca d'Italia, 2001, for a detailed analysis of this dataset). Because of the high frequencies corresponding to 0 (23.00%) or 1 (7.71%), Calabrese and Zenga (2010) suggest considering the recovery rate as a random variable with a mixed-type distribution, obtained as a mixture of a Bernoulli distribution on the set $\left\{\{0\},\{1\}\right\}$ and a (continuous) density on the interval $(0,1)$. A different approach to tackling this problem is suggested in Punzo and Zini (2010).

Attention will only be focused on the estimation of the true density of the 103511 "continuous" data on $(0,1)$. With this aim, we apply the finite mixture of beta distributions, as described in Section 2.2, with number of components g ranging from 1 to 6. Table 2 shows the values of classical model selection criteria for each value of g. The numbers in bold represent the smallest value for each row, that is, for each of them. Considering all the information criteria to hand, it make sense to choose $g = 6$ as the best "compromise" model,

Table 2: Values of AIC, BIC and CAIC for the finite mixture of beta densities with different values of g

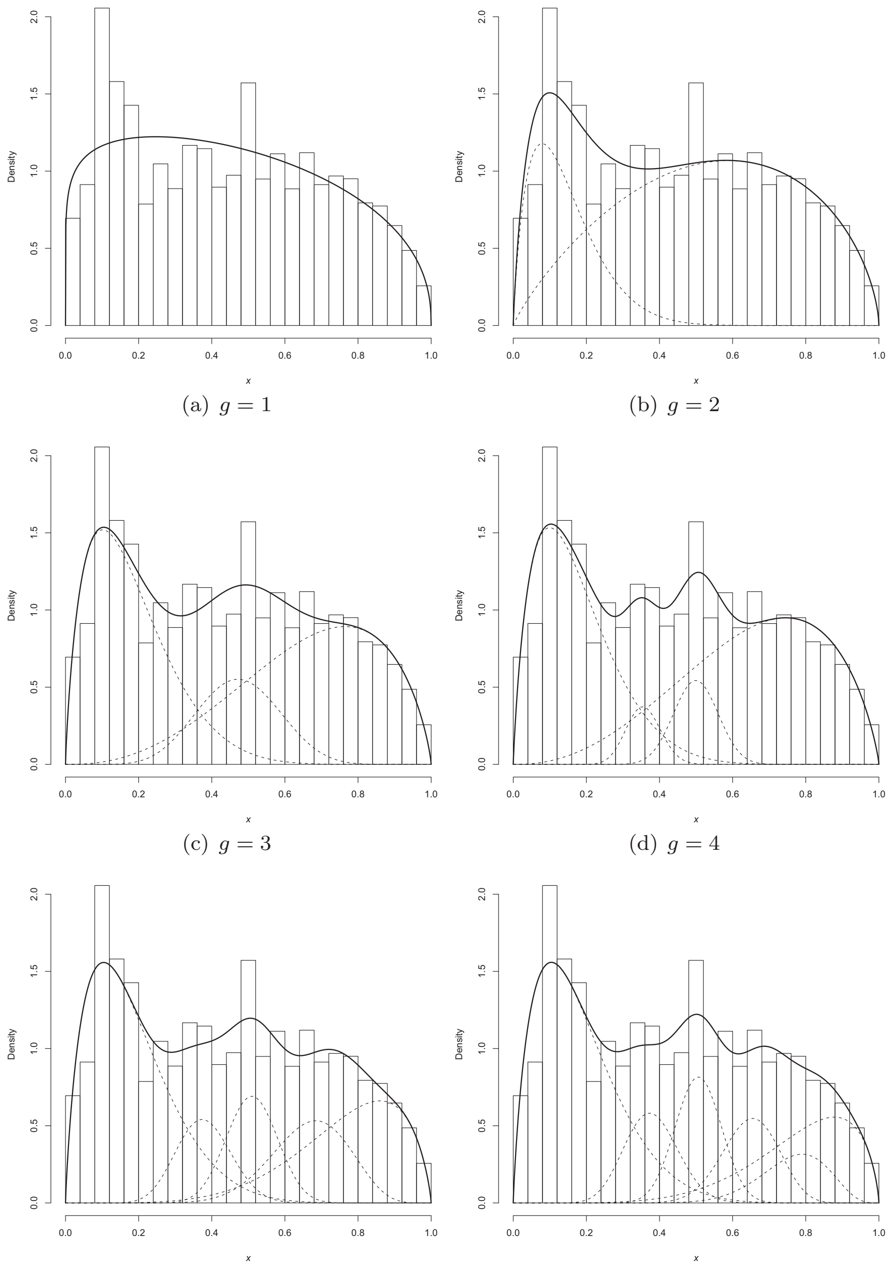
Model selection criteria	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g = 5$	$g = 6$
$-2l_c\left(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{m}}, \widehat{\boldsymbol{v}}\right)$	-7435.436	-9472.243	-9863.514	-10012.042	-10014.631	**-10158.788**
AIC	-7431.436	-9462.243	-9847.514	-9990.042	-9986.631	**-10124.788**
BIC	-7412.341	-9414.506	-9771.135	-9885.020	-9852.967	**-9962.481**
CAIC	-7410.341	-9409.506	-9763.135	-9874.020	-9838.967	**-9945.481**

Table 3: Parameters of the finite mixture of g beta densities, sorted in a non-increasing way according to the value of the modes, estimated via the BH-EM algorithm

Number of components	π_j	m_j	v_j
$g = 1$	1.000	0.246	1.782
$g = 2$	0.239	0.079	0.089
	0.760	0.588	0.669
$g = 3$	0.395	0.101	0.122
	0.156	0.471	0.055
	0.448	0.764	0.304
$g = 4$	0.381	0.099	0.113
	0.043	0.353	0.009
	0.082	0.499	0.014
	0.492	0.745	0.314
$g = 5$	0.408	0.104	0.121
	0.100	0.373	0.024
	0.120	0.509	0.020
	0.141	0.684	0.056
	0.228	0.859	0.178
$g = 6$	0.408	0.104	0.121
	0.105	0.372	0.023
	0.128	0.506	0.016
	0.105	0.654	0.027
	0.072	0.789	0.053
	0.179	0.878	0.170

corresponding to a 18-parameters model. Fig. 8 displays the fitted densities of the mixture with g ranging from 1 to 6. However, if we are interested in a more parsimonious model, $g = 4$ components could be an alternative; this choice is justified by the information criteria adopted, but also by a simple graphical inspection of Fig. 8. Details on the estimated parameters, computed via the BH-EM algorithm, are contained in Table 3.

These results confirm the observation that a single beta density (classical model used in literature for the recovery rate variable, Gupton et al, 1997; Gupton and Stein, 2002; Bruche and González-Aguado, 2010) is unable to represent X. Also, Calabrese and Zenga (2010), on the basis of their nonparametric density estimator, heuristically suggest using a mixture of a right-skewed random variable and a symmetric random variable as a (semi)parametric model for the recovery rate on bank loans (naturally, once the endpoints are removed from the support of the variable). In reality, we can see that the situation is much too complicated to be described with only two components; according to the



(a) $g = 1$

(b) $g = 2$

(c) $g = 3$

(d) $g = 4$

(e) $g = 5$

(f) $g = 6$

Fig. 8: Finite mixtures of g beta densities fitted to the observed distribution of bank loan recovery rates (Source: Bank of Italy). Dotted lines show the component densities multiplied by the corresponding weights, while the solid line plots the resulting mixture

previous remarks, a value $g \geq 4$ is required in order to obtain an adequate fit to the observed data (see Fig. 2).

Finally, it is important to underline that the parameterization in terms of the modes, immediately gives an idea of the position where the recovery rates with high probability could be located on the x-axis (see the third column of Table 3). For example, considering the three-components model, the first mode located in 0.101 suggests that a recovery rate of about 10% is more likely for this dataset.

6.2 Age at first birth in the United States

The second application is demographic. In particular, we apply a gamma mixture to the problem of modeling the distribution by age (of women) at first birth, that is, the number of first births occurring in women of a given age divided by the total number of women at first birth. To modelize such a distribution, it is well-known that in demography the use of parametric formulae is particularly emphasized, both in terms of the interpretability (substantive meaning) of the parameters and their role in facilitating spatiotemporal comparisons and forecasts (see, *e.g.*, Rogers, 1986; Congdon, 1993). The problem in this context, is that commonly used unimodal parametric distributions often do not offer adequate fit to the empirical density. An obvious example that will be analyzed, is provided by the recent behavior of multi-ethnic populations, such as in case of the the United States, where the distribution is clearly bimodal.

Motivated by these considerations, we suggest using a mixture of $g = 2$ gamma distributions defined on $S = [a, \infty)$, whose estimated modes can be used to facilitate comparisons across space and time of the two ages more representative of the distribution. Since the age is in principle a positive continuous variable, we consider $a = 0$. Naturally, as suggested by Coale (1971), it might be better to choose $a > 0$ if the values between zero and a are retained as unlikely.

We use data consisting in officially registered birth counts by calendar year, and mother's age, for the United Stated in the period 2006 (source: Human Fertility Database). Used data can be downloaded from the web site:
`http://www.humanfertility.org/cgi-bin/main.php`. Since they consist in the number n_x of first births only in correspondence to the age last birthday x, with $x = 0, 1, \ldots$, we hereafter will assume a uniform distribution of n_x within the class $[x, x + 1)$. The histogram of the distribution, with unitary bins, is displayed in Fig. 9. The gamma-mixture (solid line) is also superimposed on the plot, with dotted curves showing the component densities multiplied by the corresponding weights. Details on starting and final parameters, computed via the BH-EM algorithm, are contained in Table 4. Here, it is easy to note that the BH-initialization strategy leads the EM algorithm towards the "expected "estimates. This is particularly true in the case of the modes of the components.

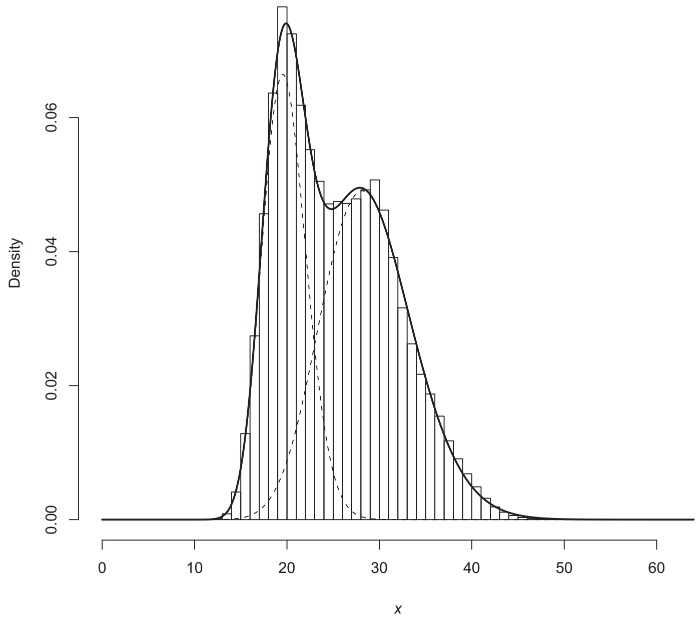


Fig. 9: Two-components gamma-mixture on $[0, \infty)$ fitted to the observed distribution, by the mother's age, of the number of first births in the United States for the period 2006 (Source: Human Fertility Database). Dotted lines show the component densities multiplied by the corresponding weights, while the solid line plots the resulting mixture

Table 4: Starting and final parameters, obtained via the BH-EM algorithm, for the mixture of $g = 2$ gamma densities on $[0, \infty)$

Parameters	π_1	m_1	m_2	v_1	v_2
Starting values	0.486	20.015	28.618	0.274	0.607
Final estimates	0.395	19.545	28.145	0.287	0.853

As can easily be seen from Fig. 9, the distribution by age at first birth gives prominence to a clear bimodality due to well-known social/demographical reasons. Roughly speaking, in the United States there is a high number of Hispanic, American Indians, Alaska Native and non-Hispanic black women younger than 20 that became pregnant, and at the same time, a high number of non-Hispanic white women having their first birth in the age range 25–29 (see Dye, 2008; Martin et al, 2009, for further details on this data set). These situations considered together, create the bimodality highlighted in Fig. 9.

7 Concluding remarks

In this paper, starting from the Pearson system of distribution, we have focused attention on the subclasses of unimodal beta and gamma distributions parameterized according to their mode. We have used these mode-parameterized densities as components inserted into a general class of finite mixtures that can be used to modelize distributions defined on bounded supports. For this mixture model, with a fixed number of components, we have described the EM algorithm for the maximum likelihood estimation of its parameters. We have also suggested an *ad hoc* initialization strategy that, in our simulation experiments, proved to be a good alternative to the classical random initialization. Real applications have also highlighted the usefulness of the parameterization applied.

As regards possible future studies, we maintain that the parameterization adopted could facilitate the analysis of modality for beta- and gamma-mixtures in line with what is usually done for finite mixtures of normal distributions (see, *e.g.*, Eisenberger, 1964; Wessels, 1964; Robertson and Fryer, 1969; Behboodian, 1970; Schilling et al, 2002; Ray and Lindsay, 2005). Moreover, focusing attention on bumps, which have been highlighted in this paper as more informative than the modes, the parameterization adopted could be useful to study the conditions leading to multy-bumpality. Finally, as a means of initializing the EM algorithm for a normal-mixture, it could also be interesting to evaluate the performance of the proposed initialization procedure, when comparing it with the k-means method.

A Details on the EM algorithm

For completeness, here we attempt to explicit the derivatives in (11) for both gamma and beta densities parameterized according to (2) and (5), respectively. We recall that the resulting ML-estimates do not have a closed-form expression and can only be computed numerically, with the aid of an iterative algorithm; such numerical methods are available in most computer software, such as `Mathematica` and `R`.

In detail, for the gamma density in (2) we have

$$\frac{\partial \ln f_j\left(x_i; m_j, v_j\right)}{\partial m_j} = \frac{1}{v_j}\left[\ln\left(x_i - a\right) - \ln v_j - \psi\!\left(\frac{m_j - a}{v_j} + 1\right)\right]$$

and

$$\begin{aligned}\frac{\partial \ln f_j\left(x_i; m_j, v_j\right)}{\partial v_j} &= \frac{1}{v_j^2}\left\{\left(m_j - a\right)\left[\ln v_j + \psi\!\left(\frac{m_j - a}{v_j} + 1\right) - \ln\left(x_i - a\right)\right] + \right.\\ &\quad \left. - \left(m_j + v_j\right) + x_i\right\}\end{aligned}$$

where $\psi\left(\cdot\right)$ is the digamma function. In the same way, for the beta density (5) results

$$\begin{aligned}\frac{\partial \ln f_j\left(x_i; m_j, v_j\right)}{\partial m_j} &= \frac{1}{v_j\left(b - a\right)}\left\{\left[\psi\!\left(\frac{b - m_j}{v_j\left(b - a\right)} + 1\right) - \psi\!\left(\frac{m_j - a}{v_j\left(b - a\right)} + 1\right)\right] + \right.\\ &\quad \left. + \ln\left(x_i - a\right) - \ln\left(b - x_i\right)\right\}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \ln f_j\left(x_i; m_j, v_j\right)}{\partial v_j} &= \frac{1}{v_j^2\left(b-a\right)}\left\{\left(b-a\right)\left[\ln\left(b-a\right)-\psi\!\left(\frac{2v_j+1}{v_j}\right)\right]+\right.\\ &+\left[\left(m_j-a\right)\psi\!\left(\frac{m_j-a}{v_j\left(b-a\right)}+1\right)+\left(b-m_j\right)\psi\!\left(\frac{b-m_j}{v_j\left(b-a\right)}+1\right)\right]+\\ &\left.-\left(m_j-a\right)\ln\left(x_i-a\right)-\left(b-m_j\right)\ln\left(b-x_i\right)\right\}.\end{aligned}$$

References

Altman E, Resti A, Sironi A (2005) Loss given default: A review of the literature. In: Altman E, Resti A, Sironi A (eds) The Next Challenge in Credit Risk Management, Riskbooks, London

Banca d'Italia (2001) Principali risultati della rilevazione sull'attività di recupero dei crediti. Bollettino di vigilanza 12

Basel Comittee on Banking Supervision (2004) International capital measurement and capital standards: a revised framework. Bank for international Settlements

Behboodian J (1970) On the modes of a mixture of two normal distributions. Technometrics 12(1):131–139

Biernacki C, Celeux G, Govaert G (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis 41:561–575

Brazier S, Sparks RSJ, Carey SN, Sigurdsson H, Westgate JA (1983) Bimodal grain size distribution and secondary thickening in air-fall ash layers. Nature 301:115–119

Bruche M, González-Aguado C (2010) Recovery rates, default probabilities, and the credit cycle. Journal of Banking & Finance 34(4):713–723

Calabrese R, Zenga M (2008) Measuring loan recovery rate: methodology and empirical evidence. Statistica & Applicazioni

Calabrese R, Zenga M (2010) Bank loan recovery rates: Measuring and nonparametric density estimation. Journal of Banking & Finance 34(5):903–911

Chen S (1999) Beta kernel estimators for density functions. Computational Statistics & Data Analysis 31:131–145

Chen S (2000) Probability density function estimation using gamma kernels. Annals of the Institute of Statistical Mathematics 52(3):471–480

Coale A (1971) Age patterns of marriage. Population studies 25(2):193–214

Congdon P (1993) Statistical graduation in local demographic analysis and projection. Journal of the Royal Statistical Society Series A Statistics in society 156(2):237–270

Cox D (1966) Notes on the analysis of mixed frequency distributions. British Journal of Mathematical and Statistical Psychology 19:39–47

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39(1):1–38

Dye JL (2008) Fertility of American Women: 2006. Current Population Reports, US Census Bureau 20(558)

Eisenberger I (1964) Genesis of bimodal distributions. Technometrics 6(4):357–363

Elderton WP, Johnson NL (1969) Systems of Frequency Curves. Cambridge University Press

Everitt B, Hand DJ (1981) Finite mixture distributions. Chapman & Hall

Gupton G, Stein R (2002) LossCalc: Moody's Model for Predicting Loss Given Default (LGD). Moody's Investors Service, New York

Gupton G, Finger C, Bhatia M (1997) CreditMetrics – technical document. J. P. Morgan & Co, New York

Izenman AJ (2008) Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer, New York

Ji Y, Wu C, Liu P, Wang J, Coombes K (2005) Applications of beta-mixture models in bioinformatics. Bioinformatics 21(9):2118–2122

Johnson NL, Kotz S (1970) Continuous Univariate Distributions, vol 1. John Wiley & Sons

Jordan MI, Xu L (1995) Convergence results for the EM approach to mixtures of experts architectures. Neural Networks 8(9):1409–1431

Kendall MG, Stuart A (1958) The Advanced Theory of Statistics, vol 1. Charles Griffin & Company Limited

Khan S, Ahmad A (2004) Cluster center initialization algorithm for K-means clustering. Pattern Recognition Letters 25(11):1293–1302

Lindsay B (1995) Mixture Models: Theory, Geometry and Applications, vol 5. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, California

Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Menacker F, Kirmeyer S, Mathews T (2009) Births: Final Data for 2006. National Vital Statistics Reports 57(7)

Mayrose I, Friedman N, Pupko T (2005) A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21(Suppl 2):151–158

Mazza A, Punzo A (2011) Discrete beta kernel graduation of age-specific demographic indicators (in press). In: Ingrassia S, Rocci R, Vichi M (eds) New Perspectives in Statistical Modeling and Data Analysis, Springer, Berlin-Heidelberg, p ..

McLachlan G, Krishnan T (2007) The EM algorithm and extensions. John Wiley & Sons, New York

McLachlan GJ, Basford KE (1988) Mixture models. Marcel Dekker

McLachlan GJ, Peel D (2000) Finite Mixture Models. John Wiley & Sons

Meilă M, Heckerman D (2001) An experimental comparison of model-based clustering methods. Machine Learning 42(1):9–29

Murphy EA (1964) One cause? Many causes? The argument from the bimodal distribution. Journal of Chronic Diseases 17(4):301–324

Pearson K (1902a) On the systematic fitting of curves to observations and measurements. Biometrika 1(3):255–303

Pearson K (1902b) On the systematic fitting of curves to observations and measurements. Biometrika 2(1):1–23

Punzo A (2010) Discrete beta-type models. In: Locarek-Junge H, Weihs C (eds) Classification as a Tool for Research, Springer, Berlin-Heidelberg, pp 253–261

Punzo A, Zini A (2010) Discrete Approximations of Continuous and Mixed Measures on a Compact Interval (in press). Statistical Papers ..(..):..–..

Ray S, Lindsay B (2005) The topography of multivariate normal mixtures. Annals of statistics 33(5):2042–2065

Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM review pp 195–239

Robertson C, Fryer J (1969) Some descriptive properties of normal mixtures. Skand Aktuarietidskr 52:137–146

Rogers A (1986) Parameterized multistate population dynamics and projections. Journal of the American Statistical Association 81(393):48–61

Schilling M, Watkins A, Watkins W (2002) Is human height bimodal? The American Statistician 56(3):223–229

Silverman B (1981) Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society, Series B: Methodological 43:97–99

Titterington DM, Smith AFM, Makov UE (1985) Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons

Wessels J (1964) Multimodality in a family of probability densities, with application to a linear mixture of two normal densities. Statistica Neerlandica 18(3):267–282