Bayesian nonparametric location–scale–shape mixtures

Noname manuscript No.
(will be inserted by the editor)

Bayesian nonparametric location-scale-shape mixtures

Antonio Canale · Bruno Scarpa

Received: date / Accepted: date

Abstract Discrete mixture models are one of the most successful approaches for density estimation. Under a Bayesian nonparametric framework, Dirichlet process location-scale mixture of Gaussian kernels is the golden standard, both having nice theoretical properties and computational tractability. In this paper we explore the use of the skew-normal kernel, which can naturally accommodate several degrees of skewness by the use of a third parameter. The choice of this kernel function allows us to formulate nonparametric location-scale-shape mixture prior with desirable theoretical properties and good performance in different applications. Efficient Gibbs sampling algorithms are also discussed and the performance of the methods are tested through simulations and applications to galaxy velocity and fertility data. Extensions to accommodate discrete data are also discussed.

Keywords Discrete random probability measures · Model-based clustering · Skew-normal distribution · Rounded mixture priors

Mathematics Subject Classification (2000) MSC 62F15 · MSC 62E15

1 Introduction

Discrete mixture models are routinely used for univariate and multivariate density estimation. A discrete mixture model characterizes the density of $y \in$

A. Canale
Department of Economics and Statistics
University of Turin and Collegio Carlo Alberto, Italy
E-mail: antonio.canale@unito.it

B. Scarpa
Department of Statistical Sciences
University of Padua, Italy
E-mail: scarpa@stat.unipd.it

$\mathcal{Y} \subset \mathbb{R}$ as

$$f(y) = \sum_{h=1}^{k} \pi_h K(y; \theta_h), \tag{1}$$

where $\sum_{h=1}^{k} \pi_h = 1$ and $K(\cdot; \theta)$ is a kernel function parametrized by a vector of parameters θ. In (1), k can be any finite integer leading to a finite mixture model, or ∞ leading to an infinite, or nonparametric, mixture model. Bayesian mixture models generalize model (1) by

$$f(y) = \int K(y; \theta) dP(\theta), \;\; P \sim \Pi,$$

where P is a random mixing probability measure, and Π is a prior over the space of mixing probability measures. The Bayesian nonparametric literature dealing with such Π has recently undergone a strong development. A rich family is represented by Gibbs-type priors (Gnedin and Pitman, 2005) which particular cases are represented by the Dirichlet process (DP) (Ferguson, 1973), the two-parameter Poisson-Dirichlet process (Perman et al., 1992), or the normalized inverse Gaussian process (Lijoi et al., 2005). All these priors are convenient choices for Π since they generate discrete probability measures almost surely and thus lead to a discrete mixture as in equation (1). In many applications the default choice is the DP prior, both for practical and historical reasons. A Dirichlet process mixture (DPM) model can be written in form (1) marginalizing out P, namely

$$f(y) = \sum_{h=1}^{\infty} \pi_h K(y; \theta_h), \qquad \theta_h \overset{iid}{\sim} P_0, \qquad \pi = \{\pi_h\} \sim \text{Stick}(\alpha) \tag{2}$$

where P_0 is a base probability measure and $\text{Stick}(\alpha)$ denotes the stick-breaking process by Sethuraman (1994) with positive scalar parameter α, i.e. the general weight π_h is obtained as

$$\pi_h = V_h \prod_{l<h} (1 - V_l), \;\; V_h \overset{iid}{\sim} \text{Be}(1, \alpha).$$

An interesting feature of finite mixture models, both for continuous and count observations, is the induced clustering structure (Fraley and Raftery, 2002), so that each component can be seen as a cluster of subjects. A common choice relies on Gaussian kernels (Lo, 1984; Escobar and West, 1995) but with this choice, it may happen that redundant mixture components with similar locations are estimated. Clearly this form of overfitting may lead to an unnecessarily complex model which is particularly unappealing if the sample size is small, and it induces a lack of interpretability due to the overlapping of similar kernels. Indeed, if the data are actually made of different sub-populations, this procedure can fail to detect the real sub-population structure, if the subpopulations distributions are not symmetric.

To deal with some of these issues, Petralia et al. (2012) propose a repulsive mixture prior which favors well separated components and can lead to more

interpretable clustering structure. Another approach consists in considering mixtures of more flexible kernels, which accounts for several degrees of skewness or kurtosis. For example, Rodríguez and Walker (2014) discuss a new family of kernels K with large support on the space of unimodal density functions. Despite the extreme flexibility obtained with the latter approach, the formulation is complex and the posterior sampling is not straightforward. In this paper, instead, we explore the use of the Azzalini (1985)'s skew-normal kernel which allows us to built a nonparametric mixture model which retains both computational tractability and good theoretical properties.

Finite mixtures of skew-normals have been already discussed in the literature both in the frequentist and Bayesian context. For example, Lin et al. (2007) discuss a finite mixture of skew-normal model. The authors propose Expectation Maximization and Gibbs Sampling algorithms for the frequentist and Bayesian estimation of the parameters, respectively. Frühwirth-Shnatter and Pyne (2010), in a fully Bayesian setting, discuss mixtures of skew-normal and skew-t, motivated by multivariate data arising from biotechnological applications. They provide an interesting discussion about the number of components, involving reversible jump Markov Chain Monte Carlo and evaluation of posterior probability via information criteria. However, from a practical point of view, it is not clear how to choose the number k of components, and in practice they fixed it *a priori*. Cavatti Vieira et al. (2013) propose a DPM of skew-normal to estimate densities, obtaining promising results on some simulation scenarios. However, no discussion on clustering is made and the DPM of skew-normal is evaluate only in terms of density estimation, no theoretical justification or results are provided, and the computations are challenging. In this paper, we discuss location-scale-shape mixture models using the skew-normal kernel motivated by the seek of meaningful clustering structure. For posterior evaluation, we propose efficient sampling algorithms, which exploit recent advances in Bayesian inference for the skew-normal model (Canale and Scarpa, 2013). In addition, we provide theoretical justification of our procedure by showing large support of the prior and proving strong posterior consistency. We also introduce a new model for probability mass function estimation exploiting the rounding procedure of Canale and Dunson (2011) with skew-normal kernels in place of classic Gaussian kernels.

The rest of the paper is organized as follows. Section 2 reviews the skew-normal distribution and formalizes location-scale-shape mixture models. Section 3 discusses some theoretical properties of the DPM of skew-normal prior. Proofs are reported in the Supplementary Materials. Section 4 gives the posterior full conditional distributions representation from which a Gibbs sampling algorithm can be obtained. In Section 5 a simulation study is carried out to show the performance of the methods in finite samples. Section 6 provides two applications and Section 7 concludes the paper.

2 Models

2.1 The skew-normal distribution

A random variable X is distributed as a skew-normal (Azzalini, 1985) with location ξ, scale ω^2 and shape λ, denoted by $X \sim SN(\xi, \omega^2, \lambda)$, if its density function is

$$f_{SN}(X; \xi, \omega^2, \lambda) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\frac{x-\xi}{\omega}\right), \tag{3}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density function and the distribution function, respectively, of a standard normal, $\xi \in \mathbb{R}$, $\omega^2 \in \mathbb{R}^+$ and $\lambda \in \mathbb{R}$. Note that for $\lambda = 0$ the SN distribution reduces to the normal $N(\xi, \omega^2)$. Let $F_{SN}(x; \xi, \omega^2, \lambda)$ be the correspondent cumulative distribution function.

The skew-normal model has several stochastic representations. Some of them are interesting since they mimic real life phenomena, and others are convenient because of their nice mathematical construction. An elegant and useful stochastic representation, for example, is obtained via convolution. If $Z \sim N(0,1)$ and $V \sim N(0,1)$, and $\delta \in (-1,1)$, then

$$X = \delta|Z| + \sqrt{1-\delta^2}V \tag{4}$$

has a skew-normal distribution $X \sim SN(0, 1, \delta/\sqrt{1-\delta^2})$. The latter representation is particularly useful if we want to simulate skew-normal random variable and, after suitable adaptation, it will be used in the Gibbs sampling algorithm of Section 4.

Frequentist estimation methods typically face difficulties with the classical parametrization of the skew-normal model reported in (3). These difficulties are intrinsically tied with the likelihood function but since the pioneering paper of Azzalini (1985), a "centered parametrization" has been adopted to bypass some of these problems. The latter is induced by the bijective map $(\mu, \sigma^2, \gamma) = \varphi(\xi, \omega^2, \lambda)$, where

$$\varphi\begin{pmatrix}\xi\\ \omega^2\\ \lambda\end{pmatrix} = \begin{pmatrix} \xi - \omega b\delta \\ \omega^2\{1-(b\delta)^2\} \\ \frac{4-\pi}{2}\left\{\frac{b\delta}{\sqrt{1-(b\delta)^2}}\right\}^3 \end{pmatrix}, \tag{5}$$

$\delta = \lambda(1-\lambda^2)^{-1/2}$, $b = \sqrt{2/\pi}$, and μ, σ^2, γ are exactly the mean, variance, and third standardized cumulant, of $X \sim SN(\xi, \omega^2, \lambda)$. The Bayesian reasoning is free from these issuses since the likelihood function is weighted by the prior distribution. For this reason, the Bayesian skew-normal literature typically deals directly with the standard parametrization, as we do in the present paper (Liseo, 1990; Liseo and Loperfido, 2006; Arellano-Valle et al., 2009; Frühwirth-Shnatter and Pyne, 2010; Cabras et al., 2012).

2.2 Mixtures of skew-normals

Assume y a continuous random variable, $y \sim f$ and $f \in \mathcal{L}$ where $\mathcal{L}$ is the space of densities with respect to the Lebesgue measure. A prior on $\mathcal{L}$, is a nonparametric mixture of skew-normal if

$$f(y) = \sum_{h=1}^{\infty} \pi_h f_{SN}(y; \xi_h, \omega_h^2, \lambda_h) \tag{6}$$

in which π_h and $(\xi_h, \omega_h^2, \lambda_h)$ are random for each $h = 1, 2, \ldots$. For simplicity henceforth we focus on a DP mixture of skew-normal, i.e. we let $\pi \sim \text{Stick}(\alpha)$, and $(\xi_h, \omega_h^2, \lambda_h) \overset{iid}{\sim} P_0$. As suggested by Escobar and West (1995) we additionally assign gamma hyperprior to α, i.e. $\alpha \sim \text{Ga}(1, 1)$, where $\text{Ga}(a, b)$ denotes the gamma distribution with mean a/b and variance a/b^2. To denote a general density from the mixture model (6) we use the notation f_{MSN}.

The choice of P_0 is very important both from the applied and theoretical point of view. P_0 is a measure over $\mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ and needs to be specified. In mixture of Gaussians models the usual choice for P_0 is normal-inverse-gamma for gaining conjugacy in the blocked Gibbs samplers. In specifying P_0 here, we want to retain computational tractability while having the possibility to include, if present, prior information. A recent proposal for the Bayesian analysis of the skew-normal model has been discussed by Canale and Scarpa (2013), showing that the prior

$$P_0(\xi, \omega^2, \lambda) = N(\xi; \xi_0, \kappa\omega^2) \times \text{Ga}(\omega^{-2}; a, b) \times SN(\lambda; \lambda_0, \psi_0, \nu_0). \tag{7}$$

leads to closed form full conditional posterior distributions whose sampling can be efficiently carried out within a Gibbs sampling scheme. See Section 4 for further details. A particular case is obtained with $\lambda_0 = 0$ and $\nu_0 = 0$, leading to

$$P_0(\xi, \omega^2, \lambda) = N(\xi; \xi_0, \kappa\omega^2) \times \text{Ga}(\omega^{-2}; a, b) \times N(\lambda; 0, \psi_0), \tag{8}$$

In this case the marginal prior for λ is a normal centered in zero with variance ψ_0. This implies that the prior expected skewness for each mixture component is zero.

However, if we are motivated by finding clustering patterns and we expect that most cluster has positive (negative) skewness, the marginal prior for λ can be the general (7). For example, consider the problem of estimating the age-specific probability of childbirth for an entire area (in Section 6.2 we provide an application of this to the city of Milan). We may think to the global distribution of the age of the mother at childbirth as a mixture of skew-normal distributions in different subpopulations with different education level or socio-economic status. Although we expect that each different subpopulation has a different behavior, e.g., developing countries typically show asymmetric to the right distribution of the age of the mother at childbirth. In this case we may want to favor this positive skewness for all mixture components, and hence we can assume that the marginal prior for λ has low prior mass on the negative semi-axis. This can be achieved, for example, with $\lambda_0 = 0$ and $\nu_0 > 0$.

2.3 Mixture of rounded skew-normals

Consider the case in which $y \in \mathcal{Y}$ is a discrete random variable with $\mathcal{Y} \subseteq \mathbb{Z}$. Let $y \sim p$ and $p \in \mathcal{C}$ where $\mathcal{C}$ is the space of the probability mass functions on the integers. Following Canale and Dunson (2011), assume that $y = h(y^*)$, where $h(\cdot)$ is a rounding function defined so that $h(y^*) = j$ if $y^* \in (a_j, a_{j+1}]$, for $j \in \mathbb{Z}$, with $a_{-\infty} < \cdots < a_0 < \cdots < a_\infty$ being an infinite sequence of pre-specified thresholds that defines a disjoint partition of $\mathbb{R}$ with $a_{-\infty} = -\infty$ and $a_\infty = \infty$. Under this setting the probability mass function p of y is $p = g(f)$, where $g(\cdot)$ is the rounding function having the simple form

$$p(j) = g(f)[j] = \int_{a_j}^{a_{j+1}} f(y^*)\mathrm{d}y^*, \quad j \in \mathbb{Z}. \tag{9}$$

A prior over $\mathcal{C}$ is obtained specifying a prior for the distribution of the latent y^*. Our proposal consists in

$$y = h(y^*), \quad y^* \sim f, \quad f(y^*) = \sum_{h=1}^{\infty} \pi_h f_{SN}(y^*; \xi_h, \omega_h^2, \lambda_h), \tag{10}$$

with $\pi \sim \text{Stick}(\alpha)$, and $(\xi_h, \omega_h^2, \lambda_h) \sim P_0$ as in Section 2.2. We call this formulation DPM of rounded skew-normal. Note that a particular case of the DPM of rounded skew-normal is obtained when y is a count variable, i.e. $\mathcal{Y} = \mathbb{N}$ and $j \in \mathbb{N}$, with the thresholds being $a_0 = -\infty < a_1 < \cdots < a_\infty = \infty$. Clearly, the properties of the prior induced on the space of probability mass functions, here described, will be largely driven by the properties of the prior on the latent space. In the next section we will study first some of the properties of model (6) and then discuss the discrete case.

3 Properties

An important property that a Bayesian nonparametric procedure should hold is the consistency in frequentist sense of the final posterior, namely if a fixed density f_0 has generated the data, the posterior should concentrates on a small neighborhood of such f_0 as the sample size increases.

We first concentrate on the properties of model (6). Large support of the prior is an important property while also having a crucial role in posterior consistency. The Kullback-Leibler (KL) support of the prior Π is the set of all f_0 such that $\Pi(\mathcal{K}_\epsilon(f_0)) > 0$, where $\mathcal{K}_\epsilon(f_0)$ is a KL ϵ-neighborhood of f_0 defined as

$$\mathcal{K}_\epsilon(f_0) = \left\{ f : \int f_0 \frac{f_0}{f} < \epsilon \right\}.$$

Wu and Ghosal (2008) proved the prior positivity of KL ϵ-neighborhoods under mild regularity conditions on f_0, for DP location-scale mixture of several kernels. Among them, the authors considered the skew-normal kernel too, assuming the shape parameter as fixed. Under the theory therein for each fixed

λ_0 we have that the prior on the space of continuous univariate densities induced via

$$f(y; P, \lambda_0) = \int f_{SN}(y; \xi, \omega^2, \lambda_0) dG(\xi, \omega^2), \quad G \sim DP(\alpha G_0),$$

has large KL support.

The next theorem, which instead is in terms of location-scale-shape mixtures prior formalizes the size of the KL support of prior (6).

Theorem 1 *Let f_0 be a density over $\mathbb{R}$ with respect to Lebesgue measure and let Π denote the prior on f induced from a location-scale-shape mixture of skew-normal kernels, i.e.*

$$f(x; P) = \int f_{SN}(x; \xi, \omega^2, \lambda) dP(\xi, \omega^2, \lambda), \quad P \sim \tilde{\Pi}. \tag{11}$$

Assume that the weak support of $\tilde{\Pi}$ contains all probability measures on $\mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ that are compactly supported and that: (i) $0 < f_0(x) < M$ for some finite constant M, (ii) $|\int f_0(x) \log f_0(x) dx| < \infty$, (iii) for $a > 0$, $\int f_0(x) \log \frac{f_0(x)}{\psi_a(x)} dx < \infty$, where $\psi_a(x) = \inf_{t \in (x-a, x+a)} f_0(t)$, and (iv) for $\eta > 0$, $\int |x|^{2(1+\eta)} f_0(x) dx < \infty$. Then f_0 is in the KL support of $\tilde{\Pi}$.

The conditions on f_0 required by Theorem 1 are the same conditions for the KL support of general location-scale mixtures and can be seen as standard regularity and tail conditions. As a corollary of Theorem 1, we give the following result which formalizes the size of the support of the prior (10). The proof follows directly from Theorem 1 of Canale and Dunson (2011) and hence is omitted.

Corollary 1 *Let p_0 be a probability mass function on $\mathbb{N}$ such that $p_0 \in g(\mathcal{L}_{\Pi^*})$ where g is the mapping function in (9), Π^* is a prior defined as in (11) and $\mathcal{L}_{\Pi^*}$ is the KL support of Π^*. Say Π the prior induced by Π^* as described in Section 2.3, then p is in the KL support of Π.*

Weak posterior consistency is a direct consequence of the large KL support of the prior thanks to the theory of Schwartz (1965). This means that as the sample size increases the posterior probability of any weak neighborhood around the true data-generating distribution f_0 converges to one with P_{f_0}-probability 1. However, strong posterior consistency is more interesting. For the discrete probability mass function case, the latter also follows directly from the large KL support property as stated in the next proposition. Its proof follows directly from Theorem 2 of Canale and Dunson (2011).

Proposition 1 *Assume we observe an iid sample $y = (y_1, \ldots, y_n)$ from p_0 satisfying the conditions of Corollary 1. For any $\epsilon > 0$, if Π is the the prior defined by (10), then the posterior $\Pi(\{p : ||p - p_0||_1 < \epsilon\} \mid y_1, \ldots, y_n) \to 1$ a.s. P_{p_0}.*

To prove strong consistency for the mixture (6), we need some further conditions on the prior. Let first $J(\delta, \mathcal{L})$ denote the L_1 metric entropy of the set $\mathcal{L}$, defined as the logarithm of the minimum integer N for which there exists $f_1, \ldots, f_N \in \mathcal{L}$ such that $\mathcal{L} \subset \bigcup_{j=1}^{N} \{f : ||f - f_j||_1 < \delta\}$. To obtain strong posterior consistency we need to define a sieve, i.e. a sequence of sets which eventually grows to cover the whole parameter space satisfying the requirements of Theorem 2 of Ghosal et al. (1999) (reported in the Supplementary Materials). That theorem basically requires that such a sieve has low entropy and high prior mass. Since our model is a generalization of the classical location-scale mixture model, the asymptotic results described in what follows are expected. However their proof is not straightforward. Indeed our parameter space is bigger than that induced by a simple location-scale mixture, and thus, it is not obvious that we can cover this parameter space with a sieve with linearly increasing entropy and (exponentially) vanishing prior probability of its complement. To construct our sieve we exploit the stick-breaking representation of the Dirichlet process following an approach first proposed by Pati et al. (2013) and adapting it to the more challenging case of skew-normal kernels. To build our sieve we first introduce the set

$$\mathcal{F}_{a,u,l,s,m} = \left\{ f_{MSN} : |\xi_h| < a, l < \omega_h < u, |\lambda_h| < s, \text{for } h \leq m, \sum_{h>m} \pi_h < \epsilon \right\} \tag{12}$$

and we formalize its size in terms of metric entropy in the next lemma.

Lemma 1 *For some $a > 0$, $u > l > 0$, and $s > 0$, the set $\mathcal{F}_{a,u,l,s,m}$ in (12) has*

$$\begin{aligned} J(\epsilon, \mathcal{F}_{a,u,l,s,m}) \leq & m \log \left\{ d_1 \left(\frac{as}{l}\right) + d_2 \left(\frac{a}{l}\right) + d_3 s \log \left(\frac{u}{l}\right) + d_4 \log \left(\frac{u}{l}\right) + s + 1 \right\} \\ & + d_3 m \log(d_4 m) \end{aligned}$$

where d_1, d_2, d_3, and d_4 are constants depending on ϵ.

To conclude this section we give our main result on consistency for the model (6) with base measure (7) which combines Theorem 2 of Ghosal et al. (1999) and Lemma 1.

Theorem 2 *Assume we observe an iid sample $y = (y_1, \ldots, y_n)$ from f_0 satisfying the conditions of Theorem 1. For any $\epsilon > 0$, if Π is the the prior defined by (6)–(7), then the posterior $\Pi(\{f : ||f - f_0||_1 < \epsilon\} \mid y_1, \ldots, y_n) \to 1$ a.s. P_{f_0}.*

Proof First define the set $\mathcal{F}_n$ as the set in (12) with $a = O(\sqrt{n})$, $s = O(\sqrt{n})$, $l = O(1/\sqrt{n})$, $u = O(\exp\{n\})$, and $m = O(n/\log(n))$. Then the proof relies on showing that $\mathcal{F}_n$ satisfies the conditions of Theorem 2 of Ghosal et al. (1999). This is obvious from the definition of P_0 in (8) and our Lemma 1.

4 Computation

A Gibbs sampler for the mixture of skew-normals can be developed generalizing the slice sampler of Kalli et al. (2011). We introduce latent $S_1, \ldots, S_n$ where $S_i = h$ if the i-th subject is drawn from the h-th mixture component. With such an approach, conditionally on S_i, each observation is drawn from a single skew-normal distribution and hence the updated of each cluster-specific set of parameters can be done easily. Consider the joint density

$$f(y_i, u_i, S_i) \propto \mathbb{1}(u_i < \pi_{S_i}) f_{SN}(y_i; \xi_h, \omega_h^2, \lambda_h),$$

then the full conditional posterior distributions for the slice variables u_i and the cluster indicators S_i are

$$u_i | y_i, S_i \sim U(0, \pi_{S_i}), \tag{13}$$

$$\text{pr}(S_i = h | u_i, y_i) \propto \mathbb{1}(h : \pi_h > u_i) f_{SN}(y_i; \xi_h, \omega_h^2, \lambda_h). \tag{14}$$

To update each cluster-specific set of parameters, the stochastic representation (4) can be used, introducing latent half-normal distributed variables $\eta_1, \ldots, \eta_n$. Conditionally on such latent variables, we can consider the generic i-th observation as being normally distributed with mean $\xi_{S_i} + \delta_{S_i}\eta_i$ and variance $(1 - \delta_{S_i}^2)\omega_{S_i}^2$, with $\delta_h = \lambda_h(1 + \lambda_h^2)^{-1/2}$. Conditionally on those η_i the observations can be seen as drawn from a suitable Gaussian distribution and this allows us to gain conjugacy for the location and scale parameters of each component of the mixture.

Finally, the distributions for the shape parameters are in closed forms and belong to the unified-skew-normal class of distribution (discussed in Arellano-Valle and Azzalini, 2006, with the acronym SUN) as discussed in Canale and Scarpa (2013). In the latter paper an efficient algorithm for sampling from SUN is proposed. The algorithm is based on the stochastic representation of a SUN distribution which is represented as a weighted sum of a Gaussian and a left truncated Gaussian distributions.

The precision parameter α can be updated as in Escobar and West (1995). The complete Gibbs sampler for model (6) is reported in Algorithm 1.

For posterior computation in the discrete case, an additional data augmentation step and a modification of step 1 are required. Indeed we first need to generate the latent continuous variable y^* and then we can continue on the lines of the Gibbs sampler for the continuous case. Algorithm 2 gives the Gibbs sampler for model (10).

5 Simulation studies

To assess the performance of the proposed approaches, we conducted a simulation study comparing our location-scale-shape mixture of skew-normal with a classic location-scale mixture of Gaussians. The methods were compared based on a Monte Carlo approximation of the mean Kullback-Leibler divergence and

Algorithm 1 Gibbs sampling for posterior simulation of model (6)

1. Sample u_i and S_i as in (13) and (14).
2. Sample α using Escobar and West (1995) given n and H, the number of occupied clusters
3. Update the stick-breaking weights using

$$V_h \sim \text{Be}\left(1 + n_h, \alpha + \sum_{l=h+1}^{H} n_l\right)$$

   where n_h is the sample size of the hth cluster.
4. Update

$$\eta_i \sim N(\delta_{S_i}(y_i^* - \xi_{S_i}), \omega_{S_i}^2(1 - \delta_{S_i}^2))$$

   where δ_h is $\lambda_h/\sqrt{\lambda_h^2 + 1}$.
5. Sample (ξ_h, ω_h) from

$$N\left(\hat{\mu}_h, \hat{\kappa}_h\omega_h^2\right) \text{I-Ga}(a + n_h/2 + 1, b + \hat{b}_h)$$

   where

$$\hat{\mu}_h = \frac{\kappa\sum_{S_i=h}(y_i - \delta_h\eta_i) + (1-\delta_h^2)\xi_0}{n_h + \kappa\omega^2(1-\delta_h^2)},\ \hat{\kappa}_h = \frac{\kappa(1-\delta_h^2)}{n_h\kappa + (1-\delta_h^2)}$$

$$\hat{b}_h = \frac{1}{2(1-\delta_h^2)}\left\{\sum_{S_i=h}\eta_i^2 - 2\delta_h\sum_{S_i=h}\eta_i(y_i - \xi_h) + \sum_{S_i=h}(y_i - \xi_h)^2 + (1-\delta_h^2)(\xi_h - \xi_0)^2\right\}.$$

6. Sample λ_h from

$$\lambda_h \sim SUN_{1,n_h+1}(\lambda_h; \lambda_0, \gamma_h, \psi_0, \Delta_h, \Gamma_h), \tag{15}$$

   where $\Delta_h = [\delta_i]_{i=1,\ldots,n_h}$ with $\delta_i = \psi_0 y_i(\psi_0^2 y_i^2 + 1)^{-1/2}$, $\gamma_h = (\Delta_{1:n_h}\lambda_0\psi_0^{-1}, 0)$, and $\Gamma_h = I - D(\Delta_h)^2 + \Delta_h\Delta_h^T$, where $D(V)$ is a diagonal matrix whose elements coincide with those of the vector V.

Algorithm 2 Gibbs sampling for posterior simulation of model (10)

0 For $i = 1, \ldots, n$, generate y_i^* from the full conditional posterior

0a Generate $u_i \sim U\left(F_{SN}(a_{y_i}; \xi_{S_i}, \omega_{S_i}, \lambda_{S_i}), F_{SN}(a_{y_i+1}; \xi_{S_i}, \omega_{S_i}, \lambda_{S_i})\right)$

0b Let $y_i^* = F_{SN}^{-1}(u_i; \xi_{S_i}, \omega_{S_i}, \lambda_{S_i})$

1b Sample u_i from and (13) and S_i from

$$\text{pr}(S_i = h|-) = \mathbb{1}(h : \pi_h > u_i)p(y_i|\xi_h, \omega_h, \lambda_h)$$

2b Continue with the Gibbs sampler (steps 2–6) for the continuous case (Algorithm 1) with y_i^* in place of y_i;

L_2 distance of the posterior mean estimate (f) from the true data generating process (f_0), defined as

$$KL(f_0, f) = \int f_0(x) \log\left(\frac{f_0(x)}{f(x)}\right) \mathrm{d}x, \;\; L_2(f_0, f) = \left(\int (f_0(x) - f(x))^2 \mathrm{d}x\right)^{1/2}. \tag{16}$$

In addition to these two indexes of goodness of fit, we report the average posterior mean number of occupied components and the average posterior mean of the DP precision parameter α. To facilitate the interpretation of the Kullback-Leibler divergence, we report the transformation proposed by McCulloch (1989). Such transformation, given by $q(KL) = (1 + (1 + e^{-2KL})^{1/2})/2$ is bounded between 0.5 and 1 and thus facilitates the interpretation of the Kullback-Leibler divergence as measure of discrepancy of f from the true f_0.

In implementing the blocked Gibbs samplers of the two models the first 1,000 iterations were discarded as a burn-in and the next 5,000 samples were used to calculate the posterior mean of the density on a fine grid of points of the domain. For our mixture of skew-normals we choose, as hyperparameters, $\xi_0 = \overline{y}$, the sample mean, and $\kappa = s^2$, the sample variance, $\lambda_0 = 0$, $\psi_0 = 10$, $\nu_0 = 0$, and $a = b = 1$. Choosing the sample mean and variance for ξ_0 and κ, respectively, can be seen as a default empirical Bayes approach (Efron and Morris, 1972) used in absence of strong prior information. Similarly, hyperparameters for the mixture of Gaussian were fixed as: the location mean $\mu_0 = \overline{y}$, the location scale $\kappa = s^2$, and the precision gamma hyperparameters both equal to 1. For the precision parameter of the DP prior we assigned a Ga(1,1) hyperprior in both cases. The values of the density for a wide variety of points of the domain were monitored to check for convergence and mixing.

Several simulations have been run under different settings obtaining similar results and, in the following, we report the results for four scenarios. Four additional scenarios related to the rounded mixture models are reported in the Supplementary Materials. The first simulation scenario assumed that the data were generated as a mixture of three Gaussians, $0.35N(-2, 1) + 0.5N(4, 2) + 0.15N(5, 2.5)$, the second scenario, as a mixture of two skew-normal, $0.65SN(0, 1, 5) + 0.35SN(4, 2, 3)$, the third as a mixture of a Gamma and a Gaussian, $0.25\text{Ga}(2, 1) + 0.75N(3, 1)$, while the last one as a simple exponential distribution with mean parameter 2. For each scenario, we generated samples of sizes $n = 100, 200, 300$ and we fit the two mixture models to 1,000 replicated data sets.

The results of the simulation are reported in Table 1. For small n the two methods have similar performance in terms of Kullback-Leibler divergence and L_2 distance from the truth but, as n increases, our location-scale-shape mixture outperforms the usual location-scale mixture. Note that in Scenario 1, i.e. a finite mixture of normals, our method is perfectly comparable with a nonparametric mixture of Gaussians in terms of distances from the truth. However, despite the substantial equality of the performances in terms of goodness of fit, our mixture of skew-normal requires on average a lower number of occupied clusters, which is a key advantage of our procedure. Indeed, only for

Table 1 Kullback-Leibler divergence (tranformed via the function q) and L_2 distance for the mean posterior density, posterior mean number of occupied cluster components and posterior mean of the DP precision parameter

		Scenario 1: *mix of normals*				Scenario 2: *mix of skew-normals*			
n	Kernel	$q(KL)$	L_2	$E(k\|-)$	$E(\alpha\|-)$	$q(KL)$	L_2	$E(k\|-)$	$E(\alpha\|-)$
100	Gaussian	0.634	0.058	4.039	0.788	0.680	0.125	3.282	0.622
	SN	0.640	0.061	2.788	0.519	0.680	0.116	3.209	0.606
200	Gaussian	0.593	0.041	3.970	0.671	0.652	0.108	3.703	0.619
	SN	0.593	0.042	2.736	0.445	0.638	0.092	3.369	0.558
300	Gaussian	0.576	0.034	3.969	0.623	0.631	0.092	4.317	0.68
	SN	0.576	0.034	2.719	0.413	0.617	0.078	3.738	0.582
		Scenario 3: *mix gamma+normal*				Scenario 4: *exponential*			
n	Kernel	$q(KL)$	L_2	$E(k\|-)$	$E(\alpha\|-)$	$q(KL)$	L_2	$E(k\|-)$	$E(\alpha\|-)$
100	Gaussian	0.639	0.066	3.989	0.775	0.786	0.205	4.617	0.905
	SN	0.647	0.067	3.556	0.681	0.785	0.201	4.664	0.917
200	Gaussian	0.601	0.048	4.405	0.748	0.761	0.187	5.374	0.927
	SN	0.604	0.048	3.791	0.635	0.752	0.177	5.364	0.926
300	Gaussian	0.584	0.039	4.601	0.728	0.750	0.180	6.061	0.980
	SN	0.585	0.039	3.938	0.615	0.740	0.169	5.878	0.950

$n = 100$, in Scenario 2 and in Scenario 4 the two methods show, on average, the same number of occupied components. These differences in terms of clustering can be appreciate also through the boxplots in Figure 1, representing the distribution of the posterior mean of the number of components for all the 1,000 samples, when $n = 300$. To conclude the discussion about the induced clustering, in Figure 2 we report the heatmaps of the posterior probability of being allocated to the same cluster for the two competing models in one sample generated from Scenario 3. These posterior probabilities are computed as the proportion of MCMC iterations, after burn-in, that two observation are allocated to the same mixture component. It is evident that our mixture of skew-normals discover two main clusters where the mixture of Gaussians discover three or four clusters. Qualitatively similar results are obtained also for the rounded mixture models reported in the Supplementary Materials.

6 Applications

6.1 Galaxy data

First we applied our modeling framework to the famous Galaxy dataset (Roeder, 1990). The dataset consists on the velocity of 82 galaxies. The histogram of the speeds reveals that the data are clearly multimodal. This feature supports the Big Bang theory, as the different modes of density can be though as clusters of galaxies moving at different speed. The data analysis was already carried out via DP mixture of Gaussians by Escobar and West (1995), and we compare their results with our mixture of skew-normal.

In implementing our blocked Gibbs sampler the first 1,000 iterations were discarded as a burn-in and the next 10,000 samples were used to calculate the posterior mean of the density on a fine grid of points of the domain. As a default non informative choice, we set the hyperparameters $\xi_0 = \overline{y}$, $\kappa = s^2$, $\lambda_0 = 0$, $\psi_0 = 10$, $\nu_0 = 0$, and $a = b = 1/2$. Since the scientific interest is galactic

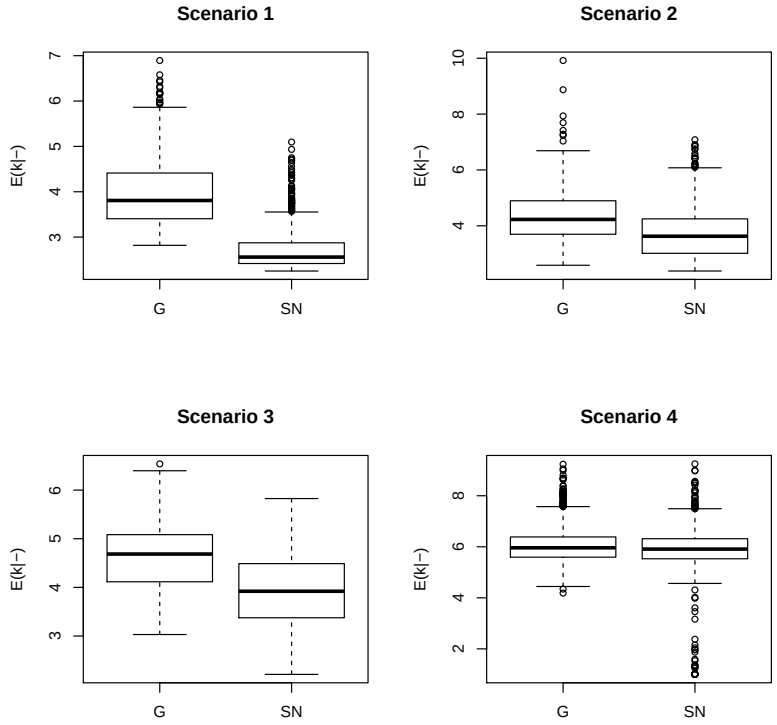
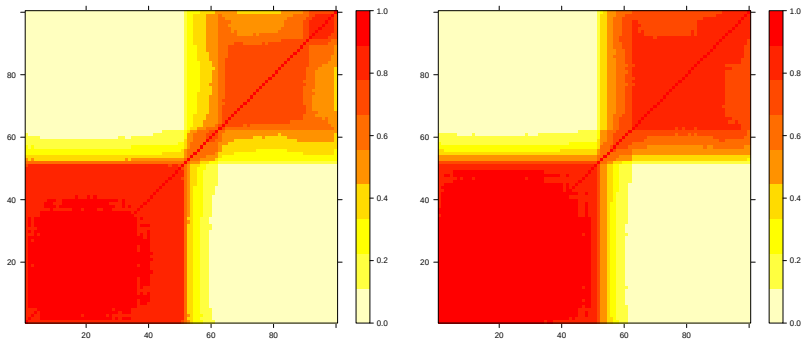


Fig. 1 Boxplot of the distributions of the posterior means of the number of components in the Gaussian mixture (G) and skew-normal mixture (SN) for all the 1,000 samples and the four scenarios and $n = 300$.

clustering, we followed Escobar and West (1995) in letting the precision DP parameter $\alpha \sim \text{Ga}(2, 4)$. The posterior mean predictive density is plotted in Figure 3 along with the relative 95% credible bands, empirical histogram, and the estimate obtained via DP mixture of Gaussians (dotted line). The two fitted densities show minor differences around the central area of the domain. Nonetheless the results in terms of density fit are similar, the posterior distribution of the number of occupied clusters and the average cluster size in the two models is different, as reported in Figure 4. From the left panel, it is evident that our approach leads to a generally lower number of occupied clusters. Our posterior distribution of the number of clusters is coherent with the posterior number of observed modes reported in Escobar and West (1995). Indeed, if a galactic cluster has a skew distribution, a single skew-normal component is sufficient, while two or more mixture components with collapsing modes are needed when using Gaussian kernels. This feature is a key advantage of using a more flexible kernel, such as the skew-normal. Such conjecture is also



(a) (b)

Fig. 2 Heatmaps of the posterior probability of being allocated to the same cluster for the mixture of Gaussians (a) and mixture of skew-normals (b) for a generic sample of size $n = 100$ generated from Scenario 3.

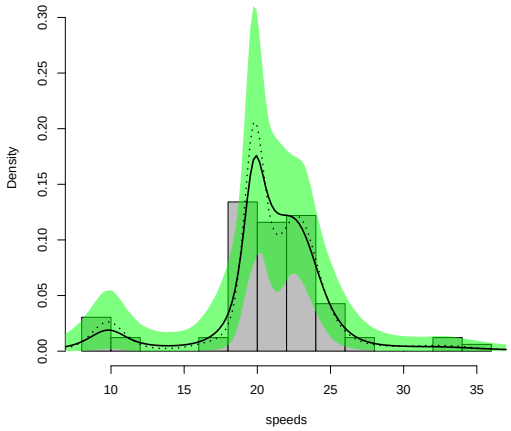


Fig. 3 Posterior estimated densities for the location-scale-shape mixture of skew-normal (continuous line) and of location-scale mixture of Gaussians (dotted line) along with the histogram of the galaxy data (number of bins calculated using Freedman and Diaconis, 1981, method). The shaded area denotes the 95% posterior confidence band for the mixture of skew-normal model.

supported by the right panel of Figure 4, that reports the posterior cluster size for both approaches. It is evident that the Gaussian mixtures needs additional clusters with few observations.

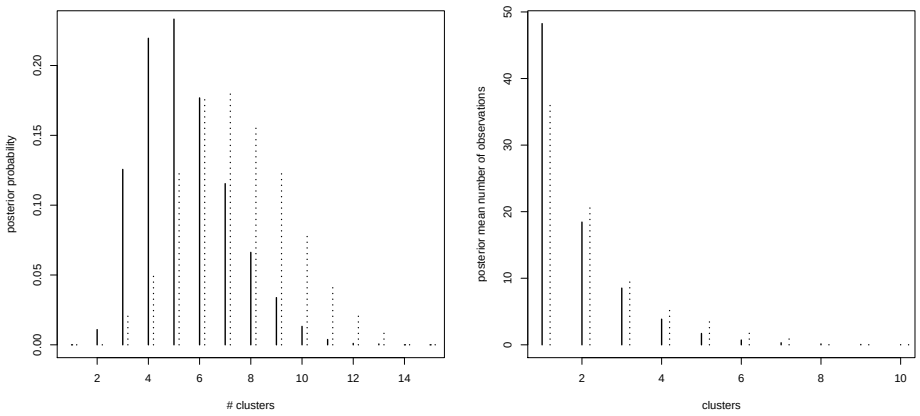


Fig. 4 Posterior probability of the number of occupied clusters (left panel) and posterior mean cluster size (right panel) in the location-scale-shape mixture of skew-normal (continuous line) and of location-scale mixture of Gaussians (dotted line) for the galaxy dataset

6.2 Childbirth age data

We apply our modeling framework to data on the births in the Milan municipality in 2011 divided by areas to estimate the different age-specific probability of childbirth. Milan is one of the biggest and multiethnic cities in Italy being the center of many economic activities and the destination of strong national and international immigration. In this context, fertility may be affected by socio-demographical and economical differences among and within the different urban areas. The presence of different subpopulations with different educational level, socio-economic status or citizenships, inside each area may give rise to asymmetric distributions of the age of the mother at childbirth. For small populations, such as the residents in Milan, there are not many specific studies on fertility indicators, and we may expect different behaviors of women with respect to the age at childbirth. Given this variety of possible patterns, a nonparametric approach to density estimation seems appropriate to both smooth the random noise affecting the curves, and to account for different patterns.

Let y be the age of the mother at childbirth and assume that we want to model the probability distribution $p(y)$. In fact, even if age is ideally continuous, data are rounded to the lower integer. Hence $p(y)$ may be seen as a

probability mass function defined on the positive integers and we estimate $p(\cdot)$ with model (10). Here, the population may be divided in a number of different sub-populations each of which may have a different fertility behavior inducing different fertility curves. Fertility curves are typically asymmetric to the right, but it is well known that more developed communities tend to postpone childhood by showing symmetric and even skew to the left fertility curves. Sub-populations living in Milan may present different patterns of fertility and a mixture model is reasonable relevant to fit available data.

Although, mixture models have been already quite used in demography, for example in the context of country age-specific fertility rate estimation where several finite mixture models have been discussed (Chandola et al., 1999; Peristera and Kostaki, 2007), to our knowledge, no skew kernel has been previously adopted. The use of the skew-normal kernel has the advantage to easily fit asymmetric pattern in each component of the mixture.

One of our goals is to compare pattern of fertility in different areas of Milan, so that we compare curves obtained by separately fitting nine models, one for each administrative zone of the Milan municipality. These nine areas include the following neighborhoods: area 1 - historical center; area 2 - central station, Gorla, Turro, Greco, Crescenzago; area 3 - Città Studi, Lambrate, Venezia; area 4 - Vittoria, Forlanini; area 5 - Vigentino, Chiaravalle, Gratosoglio; area 6 - Barona, Lorenteggio; area 7 - Baggio, De Angeli, San Siro; area 8 - Fiera, Gallaratese, San Leonardo, Quarto Oggiaro; area 9 - Garibaldi station, Niguarda.

For all zones, the hyperparameters for the base measure are set equal to $\xi_0 = \overline{y}$, $\kappa = s^2$, $\psi_0 = 1$, and $a = b = 1/2$. The DP precision parameter was assigned a Gamma hyperprior $\alpha \sim \text{Ga}(1, 1)$. The thresholds are fixed to $a_j = j$ for $j = 15, \ldots, 50$. To implement our Gibbs sampler in Algorithm 2, we discarded the first 1,000 iterations as a burn-in and we used the next 5,000 draws to calculate the posterior mean of the probability mass function for the ages $15, \ldots, 50$ years of the women. As posterior estimate, we considered the mean probability mass functions in the nine zones, reported in Figure 5 along with the empirical estimate.

Our procedure allows for smoothing across the age of childbirth and this is evident in Figure 5, where the mean of the posterior probability mass function is smoother than the empirical estimate, which has an erratic behavior, by showing a fine-scale noisy structure in the height the probability masses. However, our procedure is also able to catch the different shapes of the probability mass functions in different areas. For example, zone 1, 3, and 5 are almost symmetric with, in zone 3, only mild left skewness, and in zone 1 high concentration around the mean. These probability mass functions clearly show a delay in childbirth, with respect to classical curves, but also suggest the presence of a common fertility behavior inside these areas. These zones are in general considered quite rich neighborhoods, where lower level socio-economic family can hardly live and where immigrants are very rare: zone 1 is the center of the city, and zone 3 and 5 are the areas where typically executives and managers of the companies lives. Other areas, instead, present a small hump

around 20–25 years. In zone 4 and 6 this is fairly evident, while in zone 8 and 9 this is only partially noticeable. The former areas, where large communities of immigrants lives, are likely to have at least two subpopulations, with the smaller consisting in women anticipating the childbirth. Most of the estimated probability mass functions exhibit moderate skewness to the left, sign of a general trend of the majority of women in the area to postpone the age at childbirth, but also indicator of the presence of subgroups that anticipate it.

7 Discussion

In this paper we have discussed nonparametric location-scale-shape mixture of skew-normal kernels for density estimation and its extension to model discrete probability mass functions. For simplicity we focused on the DP prior for the mixing measures, but it is clear that the modeling framework can be generalized to general random probability measures, like the two-parameter Poisson-Dirichlet process, the normalized inverse Gaussian process. Further generalization involves substituting the univariate skew-normal kernels with their multivariate counterpart or assume a density regression setting as in Dunson et al. (2007).

The proposed location-scale-shape mixtures have the particular advantage of determining clusters with different shapes, allowing for several degrees of positive and negative skewness. This has been shown to have an impact in real applications where the model-based clustering may have some specific interpretation. We showed that this class of models has large support and asymptotic posterior consistency. Simulations confirm the asymptotic behavior and show a better quality of fit of the mixture of skew-normals with respect to the mixture of Gaussians. Evidently the number of occupied clusters is typically quite smaller in our model, thus allowing easier interpretation, when it is needed.

Acknowledgement

The authors thank Adelchi Azzalini for his shepherd into the skew world and Pierpaolo De Blasi, Stefano Mazzucco, and Igor Prünster for comments on early versions of the paper. The comments of two anonymous referees and of the Associate Editor are gratefully acknowledged.

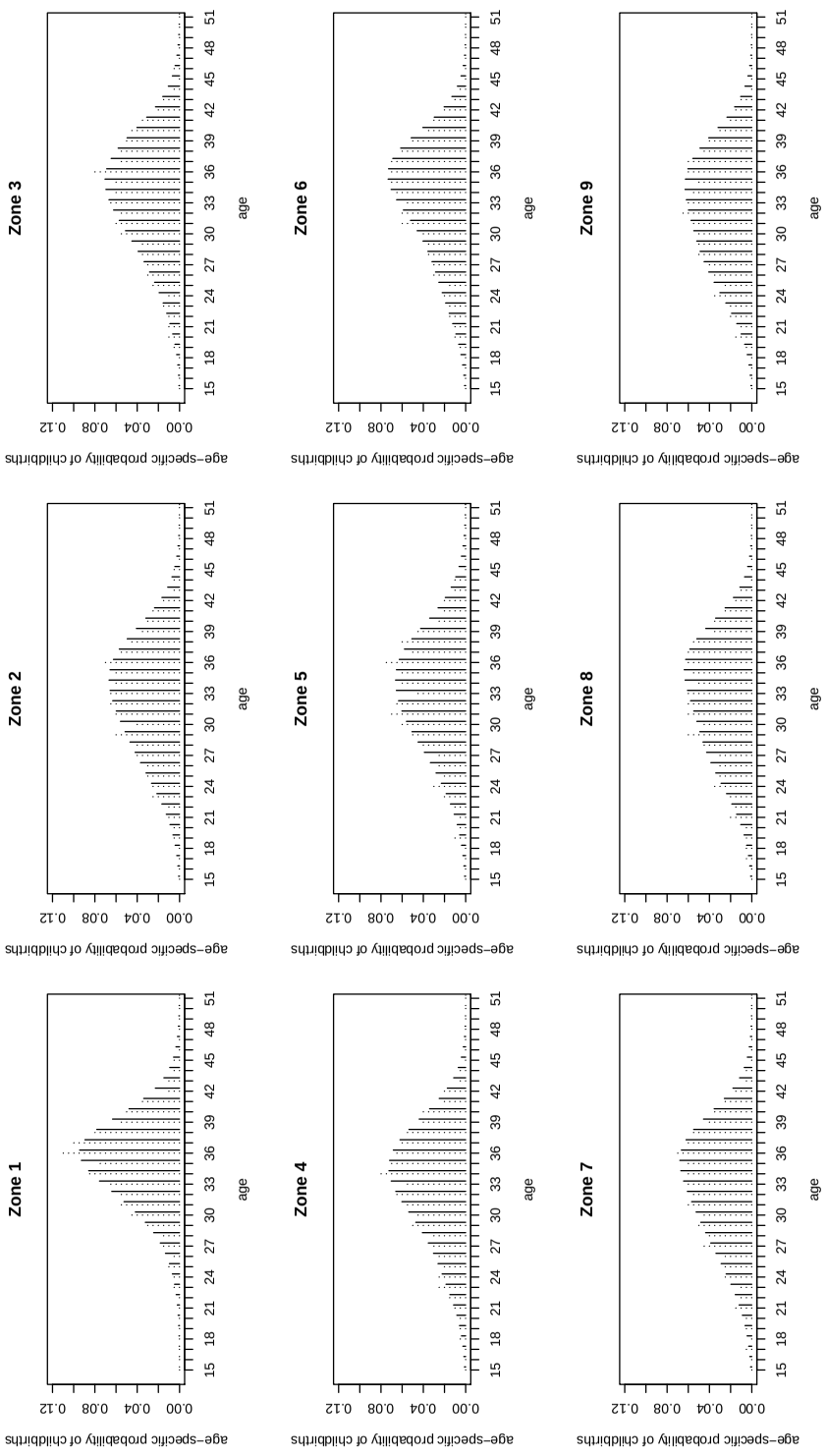


Fig. 5 Posterior mean probability mass function (black) and empirical probability mass function (dotted) for the age of the mother at childbirth in the nine zone of Milan.

References

Arellano-Valle, R. B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scand J Stat*, 33(3):561–574.

Arellano-Valle, R. B., Genton, M. G., and Loschi, R. H. (2009). Shape mixture of multivariate skew-normal distributions. *J Multivariate Anal*, 100:91–101.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand J Stat*, 12:171–178.

Cabras, S., Racugno, W., Castellanos, M., and Ventura, L. (2012). A matching prior for the shape parameter of the skew-normal distribution. *Scand J Stat*, 39:236–247.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *J Am Stat Assoc*, 106(496):1528–1539.

Canale, A. and Scarpa, B. (2013). Informative bayesian inference for the skew-normal distribution. arXiv: 1305.3080.

Cavatti Vieira, C., Loschi, R. H., and Duarte, D. (2013). Nonparametric mixtures based on skew-normal distributions: An application to density estimation. *Commun Stat A-Theor*, doi:10.1080/03610926.2013.771745:in press.

Chandola, T., Coleman, D., and Hiorns, R. W. (1999). Recent European fertility patterns: Fitting curves to 'distorted' distributions. *Pop Stud*, 53:317–329.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007). Bayesian density regression. *J Roy Stat Soc B Met*, 69:163–183.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators – part ii: The empirical Bayes case. *J Am Stat Assoc*, 67:130–139.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J Am Stat Assoc*, 90:577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann Stat*, 1:209–230.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc*, 97:611–631.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. *Probab Theory Rel*, 57:453–476.

Frühwirth-Shnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11:317–336.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann Stat*, 27(1):143–158.

Gnedin, A. and Pitman, J. (2005). Exchangeable gibbs partitions and stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325:83–102.

Kalli, M., Griffin, J., and Walker, S. (2011). Slice sampling mixture models. *Stat Comput*, 21(1):93–105.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse gaussian priors. *J Am Stat Assoc*, 100:1278–1291.

Lin, T. I., Lee, J. C., and Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Stat Sinica*, 17:909–927.

Liseo, B. (1990). La classe delle densità normali sghembe: aspetti inferenziali da un punto di vista bayesiano. *Statistica*, L:71–82.

Liseo, B. and Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution. *J Stat Plan Infer*, 136:373–389.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann Stat*, 12:351–357.

McCulloch, R. E. (1989). Local model influence. *J Am Stat Assoc*, 84:473–478.

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *J Multivariate Anal*, 116:456–472.

Peristera, P. and Kostaki, A. (2007). Modelling fertility in modern populations. *Demogr Res*, 16:141–194.

Perman, M., Pitman, J., and Yor, M. (1992). Size–biased sampling of poisson point processes and excursions. *Probab Theory Rel*, 92:21–39.

Petralia, F., Rao, V. A., and Dunson, D. B. (2012). Repulsive mixture. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *NIPS 25*, pages 1898–1906.

Rodríguez, C. E. and Walker, S. G. (2014). Univariate Bayesian nonparametric mixture modeling with unimodal kernels. *Stat Comput*, 24(1):35–49.

Roeder, K. (1990). Density estimation with confidence sets emplified by superclusters and voids in galaxies. *J Am Stat Assoc*, 85:617–624.

Schwartz, L. (1965). On Bayes procedures. *Z Wahrscheinlichkeit*, 4:10–26.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Stat Sinica*, 4:639–650.

Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron J Stat*, 2:298–331.