

Skew-normal nonparametric mixture models

Modelli mistura non parametrici di normali asimmetriche

Antonio Canale and Bruno Scarpa

Abstract In this paper we investigate nonparametric mixtures of skew-normal kernels under a Bayesian perspective. The choice of this kernel function allows us to formulate nonparametric location-scale-shape mixture priors with large support and good mathematical and computational properties. The performance of the method are tested through simulations and an application to galaxy velocity data.

Abstract *In questo articolo si studiano modelli mistura non parametrici con kernel normali asimmetrici sotto un approccio bayesiano. La scelta di questa funzione di kernel permette di formulare modelli mistura su posizione, scala e forma. Il modello proposto ha ampio supporto e ottime proprietà matematiche e computazionali. Le prestazioni del metodo sono studiate tramite studi di simulazione e un'applicazione a dati relativi alla velocità delle galassie.*

Key words: Bayesian nonparametrics; Clustering; Density estimation; Dirichlet process mixtures;

1 Introduction

Discrete mixture models are one of the most successful approaches for density estimation. A discrete mixture model assumes

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

Antonio Canale

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

Bruno Scarpa

Department of Statistics, University of Padua e-mail: scarpa@stat.unipd.it

where $y \in \mathcal{Y} \subset \mathbb{R}$. The weights of the mixture are such that $\sum_{h=1}^k \pi_h = 1$ and $K(\cdot; \theta)$ represents a kernel function parametrized by θ . Bayesian mixture models generalize model (1) by

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

where P is a mixing measure (in equation (1) this measure is discrete), and Π is a prior over the space of mixing measures. Stick breaking priors (Ishwaran and James, 2001) are convenient choices for P since a draw from a stick-breaking prior is a discrete probability measure almost surely. Among them, the most used is the Dirichlet process (DP) prior (Ferguson, 1973). A Dirichlet process mixture (DPM) model can be written similarly to (1) marginalizing out P , namely

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

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 . The choice of a Gaussian kernel $K(\cdot; \theta)$ gives the DPM of Gaussians (Lo, 1984; Escobar and West, 1995) which is computationally convenient and has nice theoretical properties. For example, it has been proved (Lo, 1984) that it can approximate any continuous density.

Finite mixture models naturally induce a clustering structure of the observations (Fraley and Raftery, 2002), so that each component can be seen as a cluster of units whose results can be eventually lead to some sort of scientific interpretation. However, a common concern is related to the number of mixture components allowed, i.e., it may happen that redundant mixture components with similar parameters are estimated.

Clearly, when the data actually show different sub-populations, the choice of Gaussian kernel leads to symmetric clusters. However, if these sub-populations are not symmetric, this procedure can fail to detect the real sub-population structure. To deal with this issue, we explore the use of the Azzalini (1985) skew-normal kernel within the nonparametric mixture model framework, which allows the model to retain both computational tractability and good theoretical properties.

The rest of the paper is organized as follows. Section 2 introduces the model and the method. Section 3 discusses a simulation study and Section 4 provides an interesting application to astronomy.

2 Model and method

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

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

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

The skew-normal model has several stochastic representations (see, e.g., Azza-
lini, 2014). 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

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

has a skew-normal distribution $SN(0, 1, \delta/\sqrt{1 - \delta^2})$.

To introduce our mixture of skew-normal kernels, assume y a continuous random
variable, $y \sim f$, with $f \in \mathcal{L}$, where $\mathcal{L}$ is the space of densities with respect to the
Lebesgue measure. A prior on $\mathcal{L}$ is a DPM of skew-normal if

$$f(y) = \sum_{h=1}^{\infty} \pi_h f_{SN}(y; \xi_h, \omega_h, \lambda_h), \quad (5)$$

with $\pi \sim \text{Stick}(\alpha)$, and $(\xi_h, \omega_h, \lambda_h) \stackrel{iid}{\sim} P_0$. To conclude the prior specification, we
may assign a gamma hyperprior to α as suggested by Escobar and West (1995).
Namely $\alpha \sim \text{Ga}(a_\alpha, b_\alpha)$, with a_α and b_α small in order to have a distribution with
heavy tails and favoring smaller values of α .

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. Under suitable assumptions on the true f_0
and the base measure P_0 , it can be shown that the prior in (5) has large support and
leads to a posterior distribution strongly consistent (Canale and Scarpa, 2013).

A Gibbs sampler for the mixture of skew-normals can be developed generalizing
the slice sampler of Kalli et al. (2011) and exploiting the stochastic representation
(4). The complete Gibbs sampler for model (5) is reported in Algorithm 1 of Canale
and Scarpa (2013) and is omitted here for brevity.

3 Simulation study

A simulation study has been conducted to assess the performance of the proposed
approach. The first simulation case assumed that the data were simulated as a mix-
ture of three Gaussians, $0.35N(-2, 1) + 0.5N(4, 2) + 0.15N(5, 2.5)$, the second sce-
nario, 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 $n = 50, 100, 200$ and we fit the two mixture models to 1,000 replicated data sets. Our method is compared with a classic location-scale mixture of Gaussians with a Monte Carlo approximation of the mean Kullback-Leibler divergence and of the L_2 distance, reported in Table 1.

Table 1 Kullback-Leibler divergence and L_2 distance for the mean posterior densities, posterior mean number of occupied cluster components and posterior mean of the DP precision parameter

n	Kernel	Scenario 1: <i>mix of normals</i>				Scenario 2: <i>mix of skew-normals</i>			
		KL	L_2	$E(k -)$	$E(\alpha -)$	KL	L_2	$E(k -)$	$E(\alpha -)$
50	Gaussian	0.237	0.146	3.940	0.908	0.391	0.280	3.137	0.694
	Skew-normal	0.312	0.158	2.800	0.614	0.465	0.286	3.241	0.720
100	Gaussian	0.107	0.099	4.039	0.788	0.261	0.242	3.282	0.622
	Skew-normal	0.118	0.103	2.788	0.519	0.260	0.225	3.209	0.606
200	Gaussian	0.051	0.070	3.970	0.671	0.182	0.209	3.703	0.619
	Skew-normal	0.051	0.071	2.736	0.445	0.148	0.178	3.369	0.558
n	Kernel	Scenario 3: <i>mix gamma+normal</i>				Scenario 4: <i>exponential</i>			
		KL	L_2	$E(k -)$	$E(\alpha -)$	KL	L_2	$E(k -)$	$E(\alpha -)$
50	Gaussian	0.260	0.165	3.482	0.789	1.237	0.497	3.884	0.884
	Skew-normal	0.308	0.163	3.331	0.749	1.284	0.495	4.161	0.955
100	Gaussian	0.140	0.124	3.989	0.775	0.967	0.455	4.617	0.905
	Skew-normal	0.159	0.125	3.556	0.681	0.959	0.445	4.664	0.917
200	Gaussian	0.073	0.089	4.405	0.748	0.781	0.414	5.374	0.927
	Skew-normal	0.078	0.090	3.791	0.635	0.720	0.392	5.364	0.926

The mixture of Gaussians fit often requires a higher number of occupied clusters. Our location-scale-shape mixture has generally comparable performances in terms of KL and L_2 distance from the true distribution. However, for small samples ($n = 50$), our proposal has higher measures of distance if compared with the mixture of Gaussians. In fact, this is not surprising, since it is well known that often the inference with the skew-normal model is not particularly efficient for small sample sizes. As expected, for high n our method is perfectly comparable and sometimes preferable, to the mixture of Gaussians.

4 Applications to Galaxy data

First we applied our modeling framework to the 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 thought 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-normals.

Since the scientific interest is galactic clustering, we follow Escobar and West (1995) in letting the precision DP parameter $\alpha \sim \text{Ga}(1/2, 50)$. The posterior mean predictive density is plotted in Figure 1 along with the empirical histogram and the estimate obtained via DP mixture of Gaussians. Our fitted density turns out to be smoother, in particular around the central area of the domain where the DPM of Gaussians clearly detects two separate modes.

The posterior distribution of the number of occupied clusters in the two models is reported in Figure 2. It is evident that our approach leads to a generally lower number of occupied clusters and that the posterior distribution of the number of clusters is coherent with the number of observed modes of the density. Indeed, if a galactic cluster is skewed, a single skew-normal component is sufficient, while two or more mixture components with collapsing modes are needed when using Gaussian kernels.

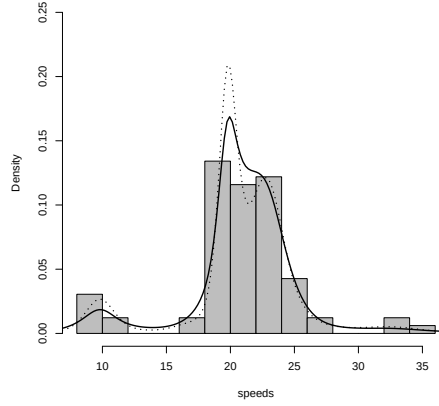


Fig. 1 Posterior estimated densities for the location-scale-shape mixture of skew-normal (continuous line) and of location-scale mixture of Gaussians (dotted line) along with the histogram of the galaxy data

5 Final remarks

In this paper we have discussed nonparametric location-scale-shape mixture of skew-normal kernels for density estimation. This class of models has the particular advantage of determining clusters with different shapes, allowing for several degrees of positive and negative skewness. This has been shown to be useful in real applications where the induced clustering may have some specific interpretation.

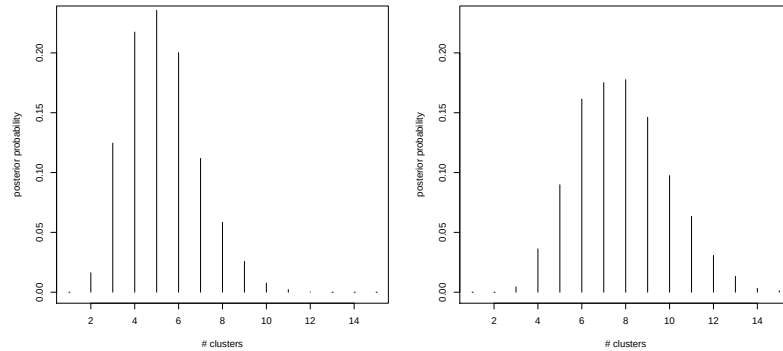


Fig. 2 Posterior probability of the average number of occupied clusters in the location-scale-shape mixture of skew-normal (left) and of location-scale mixture of Gaussians (right) for the galaxy dataset

References

- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12:171–178.
- Azzalini, A. (2014). *The Skew-Normal and Related Families*. Cambridge University Press, Cambridge.
- Canale, A. and Scarpa, B. (2013). Bayesian nonparametric location-scale-shape mixtures. arXiv:1311.7582.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. of the American Statistical Association*, 90:577–588.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230.
- Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. of the American Statistical Association*, 97:611–631.
- Ishwaran, H. and James, Lancelot, F. (2001). Gibbs sampling methods for stick breaking priors. *J. of the American Statistical Association*, 96(453):161–173.
- Kalli, M., Griffin, J., and Walker, S. (2011). Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105.
- Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12:351–357.
- Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. Statistical Association*, 85:617–624.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.