Skew mixture models for loss distributions: a Bayesian approach

Bernardi, Mauro and Maruotti, Antonello and Lea, Petrella

Sapienza University of Rome, University of Roma Tre, Sapienza University of Rome

2012

Skew Mixture Models for Loss Distributions: A Bayesian Approach

Mauro Bernardi[a], Antonello Maruotti[b, c], and Lea Petrella[a]

[a]MEMOTEF, Sapienza University of Rome, Italy
[b]DIPES, University of "Roma Tre", Italy
[c]S3RI, University of Southampton, UK

July 4, 2012

Abstract

The derivation of loss distribution from insurance data is a very interesting research topic but at the same time not an easy task. To find an analytic solution to the loss distribution may be mislading although this approach is frequently adopted in the actuarial literature. Moreover, it is well recognized that the loss distribution is strongly skewed with heavy tails and present small, medium and large size claims which hardly can be fitted by a single analytic and parametric distribution. Here we propose a finite mixture of Skew Normal distributions that provides a better characterization of insurance data. We adopt a Bayesian approach to estimate the model, providing the likelihood and the priors for the all unknow parameters; we implement an adaptive Markov Chain Monte Carlo algorithm to approximate the posterior distribution. We apply our approach to a well known Danish fire loss data and relevant risk measures, as Value-at-Risk and Expected Shortfall probability, are evaluated as well.

Keywords: Markov chain Monte Carlo, Bayesian analysis, mixture model, Skew-Normal distributions, Loss distribution, Danish data.

1 Introduction

Fitting an adequate loss distribution to real insurance data sets is a relevant problem and not an easy task in actuarial literature, mainly due to the nature of the data, which shows several features to be accounted for. In the last decades a great time and effort have been spent in this branch of research with the aim of developing always more sophisticated models, to deal with all data features. Among different approaches, the analytical method consisting in estimating the unknown parameters of a given parametric family of probability distributions has been the most adopted in the actuarial literature. Since it is well recognized that the loss distribution is strongly skewed with heavy tails, different candidates for claim severity distribution have been considered in the applications: the log-Normal, the Burr, the Weibull the Gamma and the Generalized Pareto distribution in the context of Extreme Value Theory see for example McNeil (1997), Embrechts et al. (1997) and Burnecki et al. (2010) and references cited therein. Despite their extensive application, the theoretical properties of these distributions are not always empirically matched by observed stylized facts of insurance data. Recently Bolance et al. (2008) provides strong empirical evidence in favor of the use of the Skew Normal, and log-Skew Normal distributions to model bivariate claims data from the Spanish motor insurance industry, while Ahn et al. (2012) use the log Phase-type distribution as a parametric alternative in fitting heavy tailed data. Eling (2012) shows that the Skew Normal and the Skew-Student t distributions are reasonably competitive compared to other models when describing insurance data. Unfortunately fitting Skew Normal or Skew t distributions on positive data using a frequentist approach, as in Eling (2012), can lead to unappealing estimates because of the unboundedness of the likelihood with respect to the skewness parameter. Moreover, as pointed out by Burnecki et al. (2010), usually claims distributions show the presence of small, medium and large size claims, characteristics that are hardly compatible with the choice of fitting a single parametric analytical distribution.

Because of previous considerations, in this paper we propose to use a finite mixture approach to model loss distribution data. Mixture models provide high flexibility in handling data generating process displaying several latent regimes. In addiction they are able to account for skewness and heavy tails frequently observed in actuarial data. Within the general framework of mixture models, relevant contributions to model the severity distributions have been proposed by Cooray and Amanda (2005), Scollnik (2007), who applied several composite log-Normal-Pareto models, and by Frigessi et al. (2002) who consider a dynamic mixture approach. More recently, in a different context, Sattayatham and Talangtam (2012) model motor insurance claims data from Thailand using a mixture of log-Normal distributions. Extending Eling (2012), in this paper, we propose a mixture of Skew-Normal distributions introduced by Azzalini (1985), having the following general form:

$$h(y|\theta,\eta) = \sum_{l=1}^{L} \eta_l f_{\mathsf{SN}}(y|\theta_l) \tag{1.1}$$

where L is the number of mixture components, y is the observed data, η_l are the components weights satisfying $0 < \eta_l \leq 1, \forall l = 1, 2, \ldots, L$ and $\sum_{l=1}^{L} \eta_l = 1$, while $f_{\mathsf{SN}}(y|\cdot)$ is the following

Skew Normal density function:

$$f_{\mathsf{SN}}\left(y|\xi,\omega^2,\alpha\right)=2\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left(\frac{\alpha\left(y-\xi\right)}{\omega}\right)\mathbb{I}_{(-\infty,+\infty)}\left(y\right).$$

Here $\phi()$ and $\Phi()$ stand for the density and the distribution function of a Normal random variable respectively, and $\{\xi,\omega^2,\alpha\}\subset\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}$ are the location, the scale and the skewness parameter of the Skew Normal distribution. The Skew Normal distribution is a generalization of the Gaussian law and represent a natural choice in all practical situations in which data displays skewness and kurtosis. The additional asymmetry parameter α allows for greater shape flexibility with respect to the Gaussian case achievable by setting $\alpha=0$. From a theoretical point of view the Skew Normal distributions retain much of the properties of the Gaussian distributions: in particular they are continuous unimodal random variables and their square is distributed as a Chi-square with one degree of freedom. For a complete and up to date overview of Skew Normal distributions and their applications, see the book edited by Genton (2004).

The mixture of Skew Normal distributions is a versatily tool able to capture the skewness and the kurtosis within each clustered group of claim size which is of great interest to actuaries in order to determine the appropriate level of premiums and reserves, and the re-insurance level. In this way it is possible to combine a parsimonious representation of the observed distribution with an exhaustive characterization of each group. Recently mixture of Skew Normal distribution have been proposed in different areas (see e.g. Lagona and Picone, 2012; Sahu et al., 2003).

The inferential approach we propose to fit mixture models is the Bayesian one. This approach allows to learn about the whole distribution of quantities of interest rather than just a point estimation of parameters which can be very useful in actuarial science. Moreover, when dealing with Skew Normal distributions and positive data the Bayesian paradigm overcomes the above-mentioned unbounded likelihood problem, through prior specification. Up to our knowledge this is the first attempt to consider mixture Skew Normal distributions to model severity claims in a Bayesian framework. In the following, we provide the likelihood and the priors for the all unknown parameters and we implement Markov Chain Monte Carlo (MCMC) algorithms to approximate the posterior distribution. When dealing with finite mixture models the large dimension of the parameter space and the multimodality of the posterior distribution often limit the ability of the MCMC methods to approximate the posterior distribution in a reasonable computational time. Depending on the structure of the problem and the dimension of the target distribution, even well designed algorithms could find difficulties in approximating posterior functionals, such as the mean or the mode. This is the reason why, in this paper, we do not rely on standard Metropolis-Hastings type algorithms to simulate from the posterior distribution, but we apply the new adaptive Markov chain Monte Carlo method developed in Bernardi and Petrella (2012). The algorithm combines the adaptive approach (see for example Andrieu and Thoms, 2008) with a new specification of the auxiliary distribution used to draw sample in Bayesian computation. In this way we are able to explore all the posterior modes otherwise missed by standard MCMC methods which is particular relevant when one of the objective of the inference is model selection, as it is in this paper. To this purpose we exploit the characteristics of the proposed algorithm to compute Bayes factors as strategy to choose among different mixture models. We apply our approach to the Danish fire claim data, a well known dataset investigated in actuarial literature, showing that a mixture of three Skewed components is able to capture the peculiarity of the data. Furthermore, for the chosen model, we compute the Value-at-Risk (VaR) and the Expected Shortfall probability (ES) known also as Conditional Tail Expectation; we compare our estimates with the empirical values and with estimates provided in the literature via other statistical methods. Those measures are given in a closed form formula as function of model parameters and the Bayesian approach allows to calculate their credible sets. We will show that the estimate VaR and ES perform very well at every considered confidence level.

The paper is organized as follows. Section 2 provides the basic assumption we make in a Bayesian framework as well as all the computational details needed to obtain parameter estimates. Model selection is briefly discussed in Section 2 as well. In Section 3 we analyze Danish data and provide results for model selection and parameter estimates derived from a finite mixture of Skew Normal distributions. VaR and ES are defined and estimated in Section 3 as indexes of goodness of fit. Few remarks and possible developments are discussed in Section 4.

2 Bayesian inference

Bayesian methods can be very useful in actuarial science since they enable us to learn about the whole distribution of quantities rather than just obtain point and interval estimates of each parameter. This approach assumes that all parameters in the distribution are themselves variables and that the relevant density is the posterior distribution which is proportional to the product of the model likelihood and the prior distributions.

Let $\mathsf{y} = (y_1, y_2, \ldots, y_n)$ be n independent observations drawn from the model (1.1) where we collect all the unknown parameters in the vector $\theta = \{\theta_l, \eta_l\}_{l=1,2,\ldots,L} = \{\xi_l, \omega_l^2, \alpha_l, \eta_l\}_{l=1,2,\ldots,L}$. Assuming the Skew Normal distribution for the mixture components we can write the likelihood function as:

$$L(\theta|\mathsf{y}) = \prod_{i=1}^{n} \sum_{l=1}^{L} \eta_l f_{\mathsf{SN}}(y_i|\theta_l). \tag{2.1}$$

In Bayesian setting, prior specification represents an important ingredient in developing inferential procedures. Specifying a prior distribution for finite mixture model of Skew Normal distributions entails the choice of a family of distributions for each group of component-specific parameter, location, scale and shape, $(\xi_l, \omega_l^2, \alpha_l)$, for the mixing proportions η_l and the additional elicitation of the prior hyperparameters. When dealing with finite mixture models, it is important to recognize that in order to guarantee the posterior to be a proper distribution, priors should be proper (see for example Robert, 1996, and Frühwirth-Schnatter, 2006). In addition, to prevent problems with the likelihood unbondness with respect to $\alpha_l, l = 1, 2, \ldots, L$ mentioned in the introduction and those related to model selection procedure, we avoid to be fully non informative on all parameters.

The specification of prior parameters may be particularly difficult when the parameter set is large, as in the case we consider here. For this reason, here we extend the approach proposed by Richardson and Green (1997) in the Gaussian mixtures contest, specifying the following hierarchical structure of prior distributions:

$$\begin{aligned} \xi_l &\sim \mathcal{N}\left(\zeta, \kappa^2\right) && (2.2)\\ \omega_l^2 &\sim \mathcal{IG}\left(a, b\right) && (2.3)\\ b &\sim \mathcal{G}\left(g, h\right) && (2.4)\\ \alpha_l &\sim \mathcal{T}\left(0, \varsigma^2, \frac{1}{2}\right) && (2.5)\\ \left(\eta_1, \eta_2, \ldots, \eta_L\right) &\sim \mathcal{D}\left(\delta, \delta, \ldots, \delta\right). && (2.6) \end{aligned}$$

The Gaussian prior for ξ_l is taken to be rather flat over the corresponding observed range of the data (R) as in Jasra et al. (2005); in the application considered in the next section we take $\zeta = R/2$, and $\kappa = R$. The Inverse Gamma prior for the location parameter ω_l^2 is considered introducing an additional hierarchical Gamma level prior on b the scale parameter of the distribution. To calibrate the prior hyperparameters we chose $a = 2$, $g = 0.2$ and $h = \frac{100g}{aR^2}$. For the skewness parameters α_l, as suggested by Bayes and Branco (2007), we consider a Student t distribution with one-half degrees of freedom, location and scale parameters equal to 0 and ς^2 respectively, which represents a good approximation of the Jeffreys reference prior (see e.g. Bernardo, 2005) proposed by Liseo and Loperfido (2006). Concerning the mixing proportions we apply the commonly used Dirichlet prior with $\delta > 1$, as suggested by Frühwirth-Schnatter (2006).

2.1 Bayesian Computation

During the last decades Markov Chain Monte Carlo methods, Metropolis et al. (1953) and Hastings (1970), have been extensively developed within the Bayesian approach to sample from analytically intractable posterior distributions with particular emphasis to the Gibbs Sampler and the Metropolis-Hastings algorithms. In the context of mixture models, Markov chain Monte Carlo methods have been introduced by Diebold and Robert (1994) and subsequently extended by Richardson and Green (1997) to deal with the related problem of model selection. Due to the large dimension of the parameter space and the multimodality of the posterior distributions arising in this context, standard MCMC algorithms usually fail to explore all the support of the target posterior distribution even in the simple case of Gaussian component densities, see e.g. Robert and Casella (2004), Celeux et al. (2000) and Marin et al. (2005). Moreover, when dealing with non Gaussian mixtures we face the problem of simulating from intractable full conditional densities. Recently, Frühwirth-Schnatter and Pyne (2010) proposed an equivalent stochastic representation of skewed distributions and provided a Gibbs sampler algorithm based on data augmentation for sampling the posterior parameters. Here we follow a different adaptive approach developed in Bernardi and Petrella (2012) to simulate from the posterior distribution when Skew Normal mixtures are considered. Adaptive MCMC sampling methods are simulation tools for carrying out Bayesian inference

Algorithm 2.1 Adaptive-MCMC for mixtures

1. Initialization: set $i = 0$, choose the proposal parameters $\Psi_0 = (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\nu}, \mathbf{w}_0)$, and simulate the starting values of the posterior parameters θ_0 from the prior structure defined in section 2.

2. At iteration $(i+1)$: generate a candidate draw θ^* from the proposal distribution $q_i\left(\theta, \widetilde{\Psi}, \Psi_i\right)$ defined in equation 2.13 and accept the proposed value $\theta_{i+1} = \theta^*$, with probability

$$r(\theta^*, \theta) = \min\left\{\frac{L(\theta^*|\mathsf{y})}{L(\theta|\mathsf{y})}\frac{q_i(\theta)}{q_i(\theta^*)}\frac{|J(\theta)|}{|J(\theta^*)|}, 1\right\} \tag{2.7}$$

where $|J(\theta)|$ is the determinant of the jacobian of the transformations of the parameters ω^2 and η respectively.

3. Update the proposal parameters $\Psi_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\nu}, \mathbf{w}_i)$, by the following recursions

$$\begin{aligned}
\mu_{m,i+1} &= \mu_{m,i} + \gamma_{i+1} u(m, \theta_{i+1}) w_{m,i} \bar{q}(m, \theta_{i+1})(\theta_{i+1} - \mu_{m,i}) && (2.8)\\
\Sigma_{m,i+1} &= \Sigma_{m,i} + \gamma_{i+1} u(m, \theta_{i+1}) w_{m,i} \bar{q}(m, \theta_{i+1}) \\
&\qquad \times \left[(\theta_{i+1} - \mu_{m,i})(\theta_{i+1} - \mu_{m,i})^{\mathsf{T}} - \Sigma_{m,i}\right] && (2.9)\\
w_{m,i+1} &= w_{m,i} + \gamma_{i+1} u(m, \theta_{i+1}) w_{m,i}\left[\bar{q}(m, \theta_{i+1}) - 1\right] && (2.10)
\end{aligned}$$

where $\gamma_i = \frac{1}{10 i^{0.7}}$, and

$$\begin{aligned}
u(m, \theta) &= \frac{\nu_m + d}{\nu_m + (\theta - \mu_m)^{\mathsf{T}} \Sigma_m^{-1} (\theta - \mu_m)} && (2.11)\\
\bar{q}(m, \theta) &= \frac{\mathcal{T}(\theta|\mu_m, \Sigma_m, \nu_m,)}{\sum_{m=1}^{M} w_m \mathcal{T}(\theta|\mu_m, \Sigma_m, \nu_m)}. && (2.12)
\end{aligned}$$

---

in which previous draws of the generated Markov chain are used to tailor the proposal distribution on the features of the target distribution. The distinctive characteristic of adaptive algorithms with respect to standard MCMC methods is the presence of a proposal distribution whose parameters are modified during the simulation process to minimize a distance with the target distribution.

Bernardi and Petrella (2012) propose an Independent Metropolis-Hastings sampler having the following mixture of Student t as proposal distribution:

$$q_i\left(\theta, \widetilde{\Psi}, \Psi_i,\right) = \lambda \mathcal{T}_d\left(\theta|\widetilde{\Psi}\right) + (1-\lambda)\sum_{m=1}^{M} \omega_{m,i} \mathcal{T}_d(\theta|\Psi_{m,i}) \tag{2.13}$$

where $\mathcal{T}_d(\theta|\Psi)$ denotes the probability density of a $d-$variate Student t distribution with parameters $\Psi = (\mu, \Sigma, \nu)$, $\omega_{m,i}$ are the mixture weights satisfying the constraints $0 < \omega_{m,i} \leq 1, \forall\ 1 \leq m \leq M$, and $\sum_{m=1}^{M} \omega_{m,i} = 1$, $\lambda \in (0,1)$ is the weight associated to the non-adapted distribution $\mathcal{T}_d\left(\theta|\widetilde{\Psi}\right)$, M is the number of mixture components, and the variable i controls for iterations. The presence of the fixed component guarantees the convergence of the algorithm even in cases where the parameter space is unbounded, see Andrieu and Moulines (2006) and

Haario et al. (2001), and is very useful in exploring multimodal posterior distributions when the number of modes of the proposal is misspecified. Algorithm 2.1 details the main steps of the computational method where the proposal parameters in (2.8)-(2.10) are chosen by minimizing the Kullback-Liebler divergence between the proposal and the target distribution using the stochastic approximation methods of Robbins and Monro (1951).

The algorithm allows to explore all the posterior modes, remove the data augmentation step and provide a good auxiliary distribution that can be exploited for model selection as described in the next section.

2.2 Model Selection

When different models $\{\mathcal{M}_L\}_{L=1}^{L_{\max}}$ are compatible with the data set available it is necessary to solve a model selection problem. The strategy is to calculate the Bayes factors (Kass et al., 1995 and Chipman et al., 2001), which correspond to the ratio between the marginal likelihood associated to each model i.e.

$$m\left(\mathsf{y}|\mathcal{M}_L\right) = \int L\left(\theta_L|\mathsf{y}\right) p\left(\theta_L\right) d\theta_L \tag{2.14}$$

where $L\left(\theta_L|\mathsf{y}\right)$ is the likelihood defined in equation (2.1) while $p\left(\theta_L\right)$ is the joint prior defined in (2.2)-(2.6). When a closed form for the (2.14) is not available one possibility is to approximate it by the following Importance Sampling suggested by Neal (2001):

$$m\left(\mathsf{y}|\mathcal{M}_L\right) = \frac{1}{N}\sum_{j=1}^{N} \frac{L\left(\theta_L^j|\mathsf{y}\right) p\left(\theta_L^{(j)}\right)}{h\left(\theta_L^{(j)}\right)} = \frac{1}{N}\sum_{j=1}^{N} w_j \tag{2.15}$$

where $h\left(\theta_L^{(j)}\right)$ is an importance density approximating the unnormalizing posterior distribution for the $\mathcal{M}_L$, and $\theta^{(j)}$ is the $j-$th draw from the importance density. The importance density should be carefully selected to avoid the problem of instability of the resulting estimators. We propose to use the density definined in equation (2.7) as importance density where the paramters are those of the last iteration of the MCMC algorithm. Since the proposal density considered in (2.13) guarantees the exploration of all the posterior modes, we are sufficiently comfortable of the boudness of the importance weights w_i.

3 Application to insurance claim data

The model and methodology proposed in previous sections are now applied to the insurance claim dataset of Danish fire losses. The Danish dataset consists of 2167 inflation-adjusted individual fire losses of profit (in Danish Krone, DKK) that occurred between 1980 and 1990, which can be dowloaded from `http://www.ma.hw.ac.uk/~mcneil`. We can look at Table 1 to

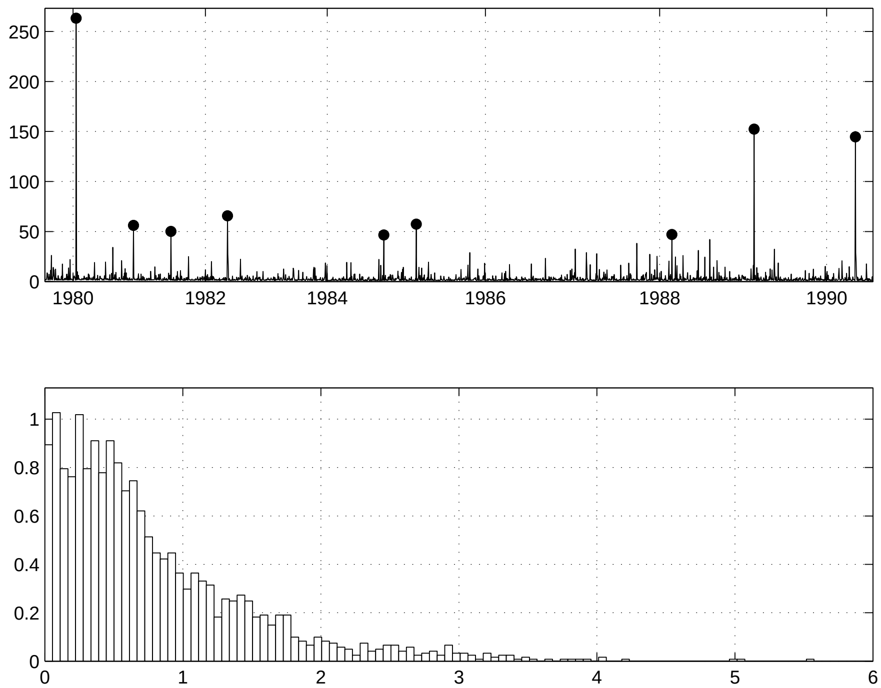

Figure 1: Time series plot (top panel) and histogram (bottom panel) of the log Danish fire losses dataset.

get more information concerning these data. Such a dataset has been extensively studied by different authors (see e.g. McNeil, 1997; Cooray and Amanda, 2005; Ahn et al., 2012). In the top panel of Figure 1 data are plotted against time showing the presence of a large amount of low-value payments followed by a small number of very large payments and their approximate times of occurrence. The heterogeneous behavior of payments made by insurance companies is one of the two main characteristics of these type of data, the other being the unpredictable nature of large losses given the past history of the series. Except during the summer periods, there is no evidence of some dynamics in the first or second moment of the series. The high kurtosis is clear from the large difference between the maximum value and the average loss (which is larger than 30 units of standard deviations), and by the fact that during the period of observations 9 values exceed 5 units of standard deviations (dark bullet in Figure 1). This is a very large number when compared with the probability of occurrence of such losses in a Gaussian world, which is less than one over 1000 years. The presence of extreme events is confirmed by the high value of the empirical VaR and ES, which are equal to 26 and 58, respectively, at the 99% confidence level. The bottom panel of Figure 1 presents the histogram of the data in logarithmic scale where it is evident the large skewness and heavy tails of the unconditional distribution, and the wide range of the data.

All previous findings as well as the presence of low, large and extreme values of loss payments supplement the various methods of checking that a different distribution is more appropriate than the Gaussian one (see e.g. Eling, 2012). Here we contribute to this branch

# of observations	2167
Mean	0.79
Standard deviation	0.72
Skewness	1.76
Kurtosis	4.18
Minimum	0.00
1st Quartile	0.28
2nd Quartile	0.56
3rd Quartile	1.09
Maximum	5.57

TABLE 1: Log Danish Data: Descriptive statistics

of research proposing a finite mixture models with skewed distributions as a flexible tool allowing for heterogeneity, skewness and kurtosis, retaining a simple interpretation of the results.

The proposed Bayesian analysis is based on the likelihood and prior specification introduced in section 2. For each mixture model we generate 25000 MCMC draws after a burn-in of 10000 and for the inference we use the adaptive MCMC algorithm specified in section 3. For the logarithm of Danish dataset we fit several finite mixture of Skew Normal (SN) distributions, and the subclass of Gaussian (N) mixtures, differing for the number of mixture components. To select the model and the number of components L we compute the marginal likelihood $m\left(\mathsf{y}|\mathcal{M}_L\right)$, for $L = 2, \ldots, L_{\max}$, as described in the previous section 2.1. Based on real-world data, the estimates of the marginal likelihood are presented in Table 2 for the Skew Normal mixtures and for the Gaussian case. In the same table we compare the results obtained with the importance sampling estimator described in section 2.2 with two differents model choice criteria, the Bayesian Information Criterion (BIC) and the Deviance Information Criteria (DIC) see Spiegelhalter et al. (2002).

The main evidence is that marginal likelihood favors the model with 3 skew components $\mathcal{M}_3^{\mathsf{SN}}$ and penalizes models with a larger number of components. The strong evidence in favor of the Skew Normal mixtures with respect to the Gaussian case is mainly due to the presence of skewness and heavy tails in the data. This conjecture is supported by the posterior estimates of the shape parameters α_l, presented in Table 3, confirming the hypothesis that the component densities are highly skewed and that Gaussian component distribution seems to be unrealistic in this context.

It is also possible to note that the BIC and the DIC select a three-component Skew Normal mixture, while looking at the Gaussian mixture models with four or five components are preferred. This result is consistent with the idea that mixtures of Gaussians usually require a large number of component densities to attain a sufficient approximation of the shape of the observed distribution. It is also important to point out that these model selection criteria do not provide consistent estimates when comparing models that are all

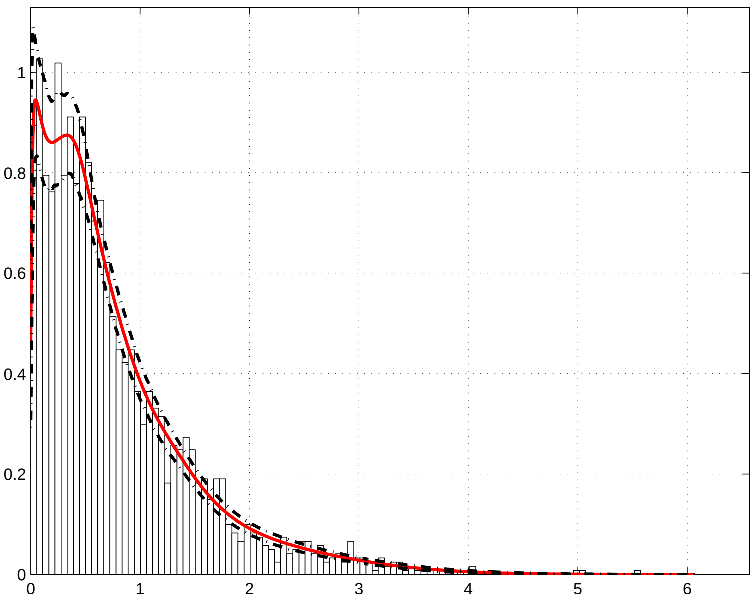

FIGURE 2: Histogram of the Danish fire data (in logarithmic scale) with superimposed density estimate of model $\mathcal{M}_3^{\mathsf{SN}}$ (red line), and 95% HPD credible set (dotted lines).

misspecified. Parameters estimate for the selected model $\mathcal{M}_3^{\mathsf{SN}}$ are presented in Table 3 which summarizes the posterior mean, the posterior median and the maximum a posteriori. The 95% high posterior density credible intervals and the Geweke's posterior convergence criteria (Geweke, 1992; 2005) are reported in the last two columns. From the results it is evident that the convergence of all the parameters have been reached. In figure 2 we superimpose the predictive distribution of the selected model $\mathcal{M}_3^{\mathsf{SN}}$ to the histogram of actual data in logarithmic scale along with the 95% credible sets. The main evidence is that tail behaviour of the observed data is well represented by the fitted distribution; in addiction with our model we are able to catch the two separated modes displayed by the observed data missed by previous approaches proposed in literature. All the point estimates denote the presence of three well separated clusters corresponding to low, medium and large losses. It could be argued that the position of the larger location parameter is far from the maximum loss, and this is due to the presence of a high peak of the data unconditonal density around zero. Posterior credible sets in the fifth column are the Bayesian analogous of the classical confidence intervals in the sense that they are used for interval estimation. As expected, the posterior credible sets for the skewness parameters are quite large and this is mainly due to the mathematical formulation of the Skew Normal density which involves the product of the Gaussian density and its cumulative density function.

Since in the actuarial practice it is important to quantify the probability of large losses, we calculate two widely used measures of risk, the Value-at-Risk (VaR) and Expected Shortfall probability (ES), at different confidence levels λ for the selected Skew Normal mixture model $\mathcal{M}_3^{\mathsf{SN}}$. The combination of the Bayesian inferential procedure with an analytical solution for

Model	n. of par	BIC/n	DIC/n	$m_{IS}(\mathsf{y}\vert\mathcal{M})$
$\mathcal{M}_2^{\mathsf{SN}}$	7	1.542	1.530	-1790.0
$\mathcal{M}_3^{\mathsf{SN}}$	11	1.549	1.523	**-1749.4**
$\mathcal{M}_4^{\mathsf{SN}}$	15	1.559	1.524	-1801.5
$\mathcal{M}_5^{\mathsf{SN}}$	19	1.573	1.518	-1824.8
Model	n. of par.	BIC/n	DIC/n	$m_{IS}(\mathsf{y}\vert\mathcal{M})$
$\mathcal{M}_2^{\mathsf{N}}$	5	1.739	1.723	-1870.6
$\mathcal{M}_3^{\mathsf{N}}$	8	1.659	1.635	**-1786.8**
$\mathcal{M}_4^{\mathsf{N}}$	11	1.599	1.565	-1803.3
$\mathcal{M}_5^{\mathsf{N}}$	14	1.586	1.590	-2155.8

Table 2: Model choice criteria for the Danish claims dataset. Skew Normal (SN) and Normal (N) mixtures. Bold faces indicates the selected model.

Parameter	Mean	Median	MaP	$\text{HPD}_{95\%}$	Geweke
ξ_1	0.006	0.006	0.003	(0.004, 0.008)	0.415
ξ_2	1.202	1.137	1.678	(0.650, 2.476)	0.500
ξ_3	0.290	0.304	0.310	(0.127, 0.388)	-0.912
ω_1^2	0.084	0.089	0.236	(0.007, 0.141)	0.656
ω_2^2	1.453	1.447	1.804	(1.039, 1.997)	0.392
ω_3^2	0.448	0.415	0.500	(0.284, 0.867)	0.337
α_1	18.215	18.139	23.092	(11.391, 25.139)	0.522
α_2	6.363	6.004	8.020	(1.132, 12.313)	-0.340
α_3	7.109	6.831	5.444	(3.677, 11.155)	-0.914
η_1	0.335	0.361	0.568	(0.107, 0.462)	0.845
η_2	0.138	0.144	0.356	(0.028, 0.213)	-0.289
η_3	0.527	0.516	0.076	(0.362, 0.736)	-1.066

Table 3: Parameter estimates obtained by fitting the Skew Normal mixture model with 3 components, $\mathcal{M}_3$ to the Danish fire loss data. Parameters are estimated by of posterior means, posterior median, and Maximum a Posteriori. The fifth column reports the 95% High Posterior Density (HPD) credible sets and the last one reports the Geweke's convergence statistics.

the risk measures as function of the Skew Normal mixture parameters allows us to provide point estimates as well as credible intervals for the VaR and ES. In particular the $\text{VaR}_\lambda(\mathsf{Y})$ at fixed λ confidence level is evaluated as the unique solution with respect to x of the following equation:

$$F_{\mathsf{Y}}^{\mathsf{SN}}(\mathsf{x},\theta) = 1-\lambda, \tag{3.1}$$

where $F_{\mathsf{Y}}^{\mathsf{SN}}(\mathsf{x},\theta)$ is the cumulative density function of the Skew Normal mixture. We compare our VaR estimates with those obtained by Eling (2012) under several (single-component)

Confidence λ	Historical	VaR_λ	Std	HPD
0.9	1.71469	1.740	0.046	(1.651, 1.831)
0.95	2.30461	2.272	0.066	(2.148, 2.403)
0.99	3.26461	3.323	0.108	(3.129, 3.552)
0.999	-	4.404	0.177	(4.073, 4.740)
0.9999	-	5.244	0.232	(4.810, 5.681)
0.99999	-	5.954	0.280	(5.435, 6.483)

TABLE 4: Value-at-Risk, VaR_λ estimates based on the MCMC output for the Danish claims dataset. Skew-Normal mixtures.

Confidence λ	Historical	ES_λ	Std	HPD
0.90	2.46989	2.404	0.101	(2.154, 2.576)
0.95	2.97039	2.891	0.103	(2.691, 3.097)
0.99	3.79335	3.805	0.136	(3.547, 4.074)
0.999	-	4.775	0.201	(4.395, 5.154)
0.9999	-	5.556	0.253	(5.089, 6.030)
0.99999	-	6.228	0.298	(5.680, 6.800)

TABLE 5: Expected Shortfall probability ES_λ estimates based on the MCMC output for the Danish claims dataset. Skew-Normal mixtures.

models. For example, the VaR at 99% confidence interval calculated in Eling (2012) with Skew-Student t distribution is 3.45, whilst we estimate a value of 3.323, which is much closer to the empirical one (equal to 3.26). In general, VaR estimates perform very well. In Table 4 results are compared with historical values for the three confidence level $\lambda = 0.9$, $\lambda = 0.95$ and $\lambda = 0.99$. VaR estimates are quite close to their empirical counterparts implying that the Skew Normal mixture model has good performance on the tail of the observed distribution. Due to lack of enough observations the same comparison could not be performed for lower confidence levels.

The ES evaluated at the VaR level x is analytically tractable for mixture of Skew-Normal, in fact it is possible to show that it is the weighted average of the Expected Shortfalls of each mixture components (Bernardi, 2012), where

$$\mathbb{ES}_{\mathsf{Y}}(\mathsf{x}, \theta) = \sum_{l=1}^{L} \pi_l \mathbb{ES}_l(\mathsf{x}, \theta_l). \tag{3.2}$$

Such a result allows us to compute the ES directly, without relying on simulations (as often pursued in the literature). The weights and the $\mathbb{ES}_l(\mathsf{x}, \theta_l)$ of each component are respectively given by the following formulas

$$\pi_l = \frac{\eta_l \left(1 - F\left(\frac{\mathsf{x}-\xi_l}{\omega_l}, \alpha_l\right)\right)}{1-\lambda}, \qquad \forall l = 1, 2, \dots, L \tag{3.3}$$

$$\mathbb{ES}_l(\mathsf{y}, \theta_l) = \xi_l + \frac{\omega_l b}{1 - F\left(\frac{\mathsf{x}-\xi_l}{\omega_l}, \alpha_l\right)} \left[\delta_l \left[1 - \Phi(\mathsf{z})\right] + \sqrt{2\pi_l}\phi_l \left(\frac{\mathsf{x}-\xi_l}{\omega_l}\right) \Phi\left(\frac{\alpha_l(\mathsf{x}-\xi_l)}{\omega_l}\right)\right], \tag{3.4}$$

with $b = \sqrt{\frac{2}{\pi_l}}$, $\delta_l = \frac{\alpha_l}{\sqrt{1+\alpha_l^2}}$ and $\mathsf{z} = \sqrt{1+\alpha_l^2}\mathsf{y}$ and $F(\cdot)$ is the distribution function of a Skew Normal variable. For more details see the analytical proofs presented in Bernardi (2012). In order to evaluate $\mathrm{VaR}_\lambda(\mathsf{Y})$ and $\mathrm{ES}_\lambda(\mathsf{Y})$ we use a Rao-Blackwellized Monte Carlo procedure using the MCMC output; this guarantees the efficiency of the estimated quantities. As for the VaR results, the ES estimates (see Table 5) provide promising results. Credible sets are increasing as the confidence level increases as expected, and in general are quite small. With a special focus on the ES estimate at the 99% confidence level, we would like to point out the goodness of the proposed approach in fitting the observed data.

4 Conclusion

In this paper we propose a mixture of Skew Normal densities for modeling the loss distribution, to deal with data displaying large and positive skewness as well as a wide right tail. Bayesian computational techniques are used to estimate parameters and for implementing the model selection procedures. Finite mixture models represent a flexible tool for fitting observed distributions allowing for heterogeneity in the underlying data while taking advantage on the simple model interpretation. These findings play a relevant role when considering extreme data with large skewness, fat tails and different latent regimes. Mixtures of Gaussian and Student t distributions, for example, are able to capture these empirical evidences providing robust counterparts to standard elliptical distributions. However, when the different underlying regimes driving the generating process exhibit a pronounced skew, more flexible component densities are required to improve the goodness of fit retaining model parsimony at the same time. The Skew Normal distribution introduced by Azzalini (1985) meets these requirements by simply adding an extra parameter controlling for the shape of distribution.

We provide empirical evidence that the chosen model with three skewed components gives a detailed description of the distribution of the modeled data. Modeling the distribution of losses is especially fruitful to predict the probability of extreme losses. Goodness of the tail approximation can be evaluated by computing two well known measures of risk, the Value-at-Risk and the Expected Shortfall probability. The availability of closed form formula for computing both measures allows us to compute efficient Rao-Blackwellized Monte Carlo estimates, marginalizing out posterior parameters uncertainty. Both measure are close to their empirical counterparts, revealing that the proposed model is able to capture the abnormal behaviour of the observed data. For the provided risk measures we also get credible regions yielding considerable insight into the uncertainty surrounding the estimates.

A straightforward extension of our proposals may concern the introduction of independent variables in a regression framework. Even if regression models for time series are widely applied in the literature, most of the provided approaches assume that all the moments of a distribution are implicitly specified through their dependence on the mean. We suggest as an interesting topic for further research, the use of regression mixtures where all the parameters of the distributions, i.e. location, scale and shape, are specified as a function of exogenous variables. This analysis may provide further insights in the analysis of the loss data where the Gaussian assumption if often violated and more complex distributions have to be taken into account.

Acknowledgments This work has been partially supported by the 2011 Sapienza University of Rome Research Project.

References

Ahn, S., Kim J. H. T., and Ramaswami, V., (2012). A new class of models for heavy tailed distributions in finance and insurance risk. *Insurance: Mathematics and Economics*, 51, 43–52.

Andrieu, C., and Moulines É., (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Annals of Applied Probability*, 16, 1462–1505.

Andrieu C. and Thoms, J., (2008). A tutorial on adaptive MCMC. *Statistical and Computing*, 18, 343–373.

Azzalini, A., (1985). A class of distributions which includes the Normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.

Bayes, C. L. and Branco, M. D., (2007). Bayesian inference for the skewness parameter of the scalar Skew Normal distribution. *Brazilian Journal of Probability and Statistics*, 21, 141–163.

Bernardi M., (2012). Risk measures for Skew-Normal mixtures. Working paper, MEMOTEF, Sapienza University of Rome.

Bernardi M., and Petrella, L., (2012). Parallel adaptive MCMC with applications. Proceedings of the 46th Italian Statistical Society Meeting.

Bernardo, J. M., (2005). Reference Analysis. In *Handbook of Statistics, 25*, Elsevier, North-Holland, Amsterdam, 459–507.

Bolance, C., Guillen, M., Pelican, E., and Vernic, R., (2008). Skewed bivariate models and nonparametric estimation for the CTE risk measure. *Insurance: Mathematics and Economics*, 43, 386–393.

Burnecki, K., Misiorek, A., and Weron, R., (2010). Loss distributions. In *Statistical Tools for Finance and Insurance*. Cizek, P., Härdle, W. K., and Weron, R. Eds. Springer-Verlag.

Celeux, G., Hurn, M. N., and Robert, C. P., (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95, 957–979.

Chipman, H., George E. I., and McCulloch E., (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes - Monograph series (2001)*, 38.

Cooray, K., and Amanda, M. A., (2005). Modeling actuarial data with a composite Lognormal Pareto model. *Scandinavian Actuarial Journal*, 5, 321–334.

Diebold, J. and Robert, C. P., (1994). Estimation of finite mixture distributions through Bayesian sampling. *Biometrika*, 56, 363–375.

Eling, M., (2012). Fitting insurance claims to skewed distributions: Are the Skew-Normal and Skew-Student good models? *Insurance: Mathematics and Economics*, 51, 239–248.

Embrechts, P., Klüppelberg, C., and Mikosch, T., (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, New York.

Frigessi, A. and Haug O. and Rue, H., (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5, 219–235.

Frühwirth-Schnatter, S., (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York.

Frühwirth-Schnatter, S., and Pyne, S., (2010). Bayesian inference for finite mixtures of univariate and multivariate Skew-Normal and Skew-t distributions. *Biostatistics*, 11, 317–36.

Genton, M. G., (2004). *Skew-Elliptical distriutions and their applications: a journey beyond normality*. Chapman and Hall.

Geweke, J., (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4*, Oxford University Press, pp. 169–193.

Geweke, J., (2005). *Contemporary Bayesian Econometrics and Statistics*. Wiley Series in Probability and Statistics. Wiley, Hoboken.

Haario H., Saksman E. and Tamminen J., (2001). An Adaptive Metropolis algorithm. *Bernoulli*, 14, 223–242.

Hastings W. K., (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

Jasra, A., Holmes, C. C., and Stephens, D. A., (2005), Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling, *Statistical Science*, 20, 50–67.

Kass, R. E., and Raftery A. E., (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Liseo, B., and Loperfido, M. A., (2006). A note on reference priors for the scalar Skew Normal distribution. *Journal of Statistical Planning and Inference*, 136, 373–389.

Lagona, F., and Picone, M., (2012). Model-based clustering of multivariate skew data with circular components and missing values, *Journal of Applied Statistics*, 39, 927–945.

Marin, J. M., Mengersen, K., Robert, C. P., (2005). Bayesian modelling and Inference on mixtures of distributions. In *Handbook of Statistics, 25*, Elsevier, North-Holland, Amsterdam, 459–507.

McNeil, A. J., (1997). Estimating the Tails of Loss Severity Distributions using Extreme Value Theory, *Astin Bulletin*, 27, 117–137.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1091.

Neal R. M., (2001). Annealed importance sampling. *Statistics and Computing*, 11, 125–139.

Richardson, S., and Green P. J., (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society Series B*, 59, 731–758.

Robert, C. P., (1996), Mixtures of distributions: Inference and estimation. In *Markov Chain Monte Carlo in Practice*. Gilks, W. R., Richardson S., and Spiegelhalter D. J. Eds. Chapman and Hall.

Robert, C. P., and Casella, G., (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York.

Robbins H. and Monro S., (1951). A stochastic approximation method. *Annals of Mathematical Stastistics*, 22, 400–407.

Sahu, S. K., Dey, D. K., and Branco, M. D., (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*, 31, 129–150.

Sattayatham P. and Talangtam T., (2012). Fitting of finite mixture distributions to motor insurance claims. *Journal of Mathematics and Stastics*, 8, 49–56.

Scollnik, D. P. M., (2007). On composite lognormal-Pareto models. *Scandinavian Actuarial Journal*, 1, 20–33.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Stastical Society, Series B*, 59, 731–792.