Insurance: Mathematics and Economics 70 (2016) 387–396

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Modeling loss data using mixtures of distributions

Tatjana Miljkovic [a,*], Bettina Grün [b,1]

[a] Department of Statistics, Miami University, 319 Upham Hall, 100 Bishop Circle, Oxford, OH 45056-1879, United States
[b] Department of Applied Statistics, Johannes Kepler Universität Linz, Altenbergerstraße 69, 4040 Linz, Austria

ARTICLE INFO

Article history:
Received December 2015
Received in revised form
May 2016
Accepted 30 June 2016
Available online 16 July 2016

JEL classification:
C02
C40
C60

Keywords:
Mixtures
Non-Gaussian distributions
EM algorithm
Risk measures
Danish Fire insurance losses

ABSTRACT

In this paper, we propose an alternative approach for flexible modeling of heavy tailed, skewed insurance loss data exhibiting multimodality, such as the well-known data set on Danish Fire losses. Our approach is based on finite mixture models of univariate distributions where all K components of the mixture are assumed to be from the same parametric family. Six models are developed with components from parametric, non-Gaussian families of distributions previously used in actuarial modeling: Burr, Gamma, Inverse Burr, Inverse Gaussian, Log-normal, and Weibull. Some of these component distributions are already alone suitable to model data with heavy tails, but do not cover the case of multimodality. Estimation of the models with a fixed number of components K is proposed based on the EM algorithm using three different initialization strategies: distance-based, k-means, and random initialization. Model selection is possible using information criteria, and the fitted models can be used to estimate risk measures for the data, such as VaR and TVaR. The results of the mixture models are compared to the composite Weibull models considered in recent literature as the best models for modeling Danish Fire insurance losses. The results of this paper provide new valuable tools in the area of insurance loss modeling and risk evaluation.

1. Introduction

Modeling insurance losses is more art than science. Techniques that sometimes work well for one data set may not be applicable to another data set. An actuary needs to weigh many factors surrounding the modeling such as risk management and pricing decisions or impact on capital requirements. Recent literature on the modeling of heavy tailed insurance loss data tends to focus more on simple models based on single parametric distributions and composite models (Bakar et al., 2015). Composite modeling is also referred to as splicing (see Klugman et al., 2012). For these models, estimation tools are in general already available, e.g., in the open-source environment for statistical computing and graphics R (R Core Team, 2015).

Limited literature exists on modeling insurance losses using K-component finite mixture models from parametric, non-Gaussian families of distributions exploring effective computational strategies. Notable exceptions are Lee and Lin (2010) and Verbelen et al. (2015, 2016) who consider finite mixtures of Erlang distributions. In this paper, we present the flexible finite mixture approach for modeling insurance losses using suitable parametric distributions, other than Erlang, for the components focusing on distributions previously proposed in the actuarial science. We show how the estimation with the expectation–maximization (EM) algorithm and model selection can be performed, and illustrate the results of this approach when applied to the well-known data set of Danish Fire losses. The Danish Fire data set is characterized as being heavy-tailed by Resnick (1997) and McNeil (1997). These authors developed several statistical plotting tools such as mean excess plot, QQ-plots, and the Hill plot for accessing the tail behavior of Danish Fire losses. These tools are available as part of the R package **evir** (Pfaff and McNeil, 2012).

The insurance losses coming from different sources are heterogeneous as reflected in multimodality, skewness, and heavy tail distributions. Mixture models can be used to capture the heterogeneity in the data and allow for the mixture components to represent groups in the population. Given the different risk assigned to each of the groups, augmenting the mixture model with a concomitant model for the weights (Dayton and Macready, 1988) would allow classifying observations into these groups and thus enable an improved risk evaluation. For these reasons, modeling the insurance losses using K-component finite mixture models is an appealing approach. In particular, the K-component finite mixture models also allow for the flexibility to easily add

* Corresponding author. Fax: +1 513 529 0989.
E-mail addresses: miljkot@miamioh.edu (T. Miljkovic), Bettina.Gruen@jku.at (B. Grün).
[1] Fax: +43 732 2468 6800.

http://dx.doi.org/10.1016/j.insmatheco.2016.06.019
0167-6687/

additional components as compared to composite modeling that is limited to two distributions only. Our modeling approach based on mixtures is contrasted with the approach proposed in the recently published paper by Bakar et al. (2015) based on composite Weibull models, which so far was found to perform best for the Danish Fire losses data set.

Different types of mixture models have been considered in the literature. Keatinge (1999) proposed modeling losses with a mixture of exponential distributions using maximum likelihood (ML) estimation based on the Newton's algorithm. While this model is useful in some actuarial applications, the mode of this model is at zero and the distribution is completely monotonic (see Wang et al., 2006), which may result in a poor fit in the case of modeling heavy-tail losses. Klugman and Rioux (2006) tried to address this issue by proposing a flexible mixture model that will include not only exponential components but also Gamma, Log-normal and Pareto components with non-negative weights that sum to one, with the restriction that either weight associated with the Gamma or Log-normal component equals zero. While this model allows for the existence of an interior mode with the inclusion of a Gamma or Log-normal component, the number of modes is still limited to at most three.

Lee and Lin (2010) proposed modeling and evaluating insurance losses via mixtures of Erlang distributions using the EM algorithm for estimation. The components in the mixture from the Erlang family were restricted to a common scale parameter to ease estimation because it allows for an effective initialization of the EM algorithm based on Tijms (1994) approximation. This restriction was justified because this class is already dense in the space of positive continuous distributions. However, it can be assumed that restricting the scale parameter leads to mixtures containing more components in order to achieve a suitable fit than would be necessary in an unrestricted setting. Lee and Lin (2010) showed that Log-normal, Gamma, and Generalized-Pareto densities can be suitably approximated with these Erlang mixtures, and they also demonstrated their proposed approach on catastrophic loss data from the United States. Verbelen et al. (2015) further extended the approach of fitting mixtures of Erlang distributions with the EM algorithm to censored and truncated data, using also the approximation by Tijms (1994) to initialize the EM algorithm. Multivariate Erlang mixtures with a common scale parameter are studied by Verbelen et al. (2016). They introduced a computationally efficient initialization and adjustment strategy iteratively used by the EM algorithm for the estimation of the shape parameter vectors, and their implementation of the EM algorithm is publicly available in the form of R code.

We extend mixture modeling beyond the Erlang family for the components and without imposing a restriction on any of the parameters. Six finite mixture models are developed with component-specific distributions from parametric, non-Gaussian families: Burr, Gamma, Inverse Burr, Inverse Gaussian, Log-normal, and Weibull. Estimation of all these models is possible using the EM algorithm, and we consider three different initialization strategies for the EM algorithm: distance-based, k-means, and random initialization. We compare our results to the composite models previously fitted to the same data sets and shown to perform best on this data set by Bakar et al. (2015). Those models use the Weibull distribution up to a threshold and a family of transformed Beta distributions beyond the threshold for modeling the heavy tail. Bakar et al. (2015) showed that composite models based on Burr, Paralogistic, and Logistic distributions for the tail fitted the real data better than those composite models based on Log-normal, Pareto (Inverse Pareto), and Gamma distributions. When comparing our results to those published by Bakar et al. (2015) using the same real data set, we show that finite mixture models may fit the data better than composite Weibull models, if the component-specific parametric family is suitably chosen.

In Section 2, we introduce the models, describe the EM algorithm for estimation, along with different initialization methods and computational strategies, propose suitable model selection criteria, and outline how risk measures can be calculated for these models. In Section 3, we apply our methodology by fitting the finite mixtures with component distributions from the six different parametric families to the well-known Danish Fire losses and discuss our findings. In the same section, we provide the results of the simulation studies. Section 4 concludes.

2. Methodology

2.1. Problem setting

Let $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ be a sample of independent and identically distributed random variables from a K-component finite mixture of probability distributions. The mixture model in parametric form is defined as

$$f(x|\boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \phi_k(x|\boldsymbol{\theta}_k), \tag{2.1}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\pi}', \boldsymbol{\theta}')' = (\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_{K-1}, \boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \ldots, \boldsymbol{\theta}_k', \ldots, \boldsymbol{\theta}_K')'$ is the vector of unknown parameters, π_k denotes the component weight of the kth component satisfying $0 < \pi_k \leq 1$, $\forall k \in \{1, \ldots, K\}$ and $\sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\theta}_k$ are the parameters of the kth density function $\phi_k(\cdot)$. We assume that the ϕ_k are density functions that are absolutely continuous with respect to the Lebesgue measure and are elements from the same univariate parametric family with a d-dimensional parameter vector $\boldsymbol{\theta}_k$, $\Im = \{\phi_k(\cdot|\boldsymbol{\theta}_k), \boldsymbol{\theta}_k \in \Theta \subset \mathbb{R}^d\}$. For a mixture as given in Eq. (2.1), the component densities $\phi_k(\cdot)$ are assumed to be from the same parametric family and differ only in component parameters $\boldsymbol{\theta}_k$. Six different density functions are considered: Burr, Gamma, Inverse Burr, Inverse Gaussian, Log-normal, and Weibull. These parametric distributions are commonly employed in modeling loss data and are thus used as basic building blocks to generate more flexible distributions by incorporating them into the finite mixture framework. Finite mixture distributions are well known for their flexibility in modeling heterogeneous data.

For estimating these finite mixture models, first ML estimates of the parameters can be obtained for a given K and parametric family using the EM algorithm as proposed by Dempster et al. (1977) and outlined in Section 2.2. Details regarding initialization of the EM algorithm and computational strategies are described in Sections 2.3 and 2.4. Then a suitable model can be selected based on model selection criteria (see Section 2.5).

2.2. The EM algorithm and parameter estimation

The EM algorithm is an iterative method for finding the ML parameter estimates of a given model and usually is employed when the data is incomplete or has missing values. The method exploits the fact that in general the maximization problem is easier for the complete data than the incomplete data. Every iteration of the EM algorithm consists of two steps: expectation (E-step) and maximization (M-step).

In the finite mixture framework, the missing observations correspond to the component identifiers. The density function $f(x|\boldsymbol{\Psi})$ in Eq. (2.1) is referred to as the incomplete data density with the associated log-likelihood $\ell_x(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\Psi})$.

For the implementation of the EM algorithm, the complete data log-likelihood function is required. We consider a random vector of complete information $\boldsymbol{C} = (\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{X}$ represents a random variable corresponding to the observed sample and $\boldsymbol{Z} = (Z_{ik} \in \{0, 1\}, i = 1, \ldots, n, k = 1, \ldots, K)$ is the set of latent random

variables indicating for all observations from which component they are. The complete data likelihood is defined as

$$L_c(\boldsymbol{\Psi}) = \prod_{i=1}^{n}\prod_{k=1}^{K}(\pi_k\phi_k(x_i|\boldsymbol{\theta}_k))^{z_{ik}}, \tag{2.2}$$

where $z_{ik} = 1$ indicates that observation x_i originated from component k; otherwise, $z_{ik} = 0$. The logarithm of (2.2) is defined as the complete data log-likelihood

$$\ell_c(\boldsymbol{\Psi}) = \log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}[\log(\pi_k) + \log(\phi_k(x_i|\boldsymbol{\theta}_k))]. \tag{2.3}$$

The E-step of the sth iteration consists of determining the conditional expectation of (2.3) given the observed data and the current parameter estimates at iteration $s-1$. This means that the expectation is taken with respect to the conditional posterior distribution of the latent data, z_{ik}, $i = 1, \ldots, n$ and $k = 1, \ldots, K$, given the observed data x_i, $i = 1, \ldots, n$ and the current parameter estimates $\boldsymbol{\Psi}^{(s-1)}$. Because the complete data log-likelihood is linear in the latent data and the linearity of taking expectations, the expected complete data log-likelihood is obtained by replacing the missing data with their expected values conditional on the observed data and the current parameter estimates. These expected values are given by

$$\pi_{ik}{}^{(s)} = \mathbb{E}[z_{ik}|x_i, \boldsymbol{\Psi}^{(s-1)}] = \frac{\pi_k{}^{(s-1)}\phi_k\left(x_i|\boldsymbol{\theta}_k{}^{(s-1)}\right)}{\sum_{k'=1}^{K}\pi_{k'}{}^{(s-1)}\phi_{k'}\left(x_i|\boldsymbol{\theta}_{k'}^{(s-1)}\right)},$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. $\pi_{ik}{}^{(s)}$ corresponds to the posterior probability that x_i comes from the kth mixture component, calculated at the sth iteration of the EM algorithm. The result is referred to as Q-function given by

$$Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}{}^{(s)}[\log(\pi_k) + \log(\phi_k(x_i|\boldsymbol{\theta}_k))].$$

Since the latent data only relate to $\boldsymbol{Z}$, the E-step does not depend on the form of density in $\Im$; thus, it is the same for all six distributions considered in this paper except for the density evaluations required.

In the M-step, new estimates for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are obtained by maximizing the Q-function. This optimization problem can be solved separately for $\boldsymbol{\pi}$ and for the parameter vector of each of the components $\boldsymbol{\theta}_k$, thus leading to several different optimization problems which, however, can be solved in closed form or at least are rather low-dimensional.

The estimates of $\boldsymbol{\pi}$ are updated in the sth iteration by

$$\hat{\pi}_k^{(s)} = \frac{1}{n}\sum_{i=1}^{n}\pi_{ik}^{(s)}.$$

New estimates for $\boldsymbol{\theta}_k$ are obtained by solving a weighted ML estimation problem for each of the different component distributions, where the weights correspond to the a-posteriori probabilities. When this weighted ML estimation problem can be solved analytically in closed form, this step is in general easy to accomplish. Alternatively, numerical optimization methods can be used to maximize the weighted log-likelihood. The following formulas indicate how the component-specific parameter estimates $\boldsymbol{\theta}_k$ can be obtained in the M-step for the different distributions considered in this paper.

Burr: $X \sim Burr(\alpha, \theta, \gamma)$
The density function of the Burr distribution with two shape parameters, α and γ, and a scale parameter θ, is given by

$$f(x|\alpha, \theta, \gamma) = \frac{\alpha\gamma\left(\frac{x}{\theta}\right)^{\gamma}}{x\left(1+\left(\frac{x}{\theta}\right)^{\gamma}\right)^{\alpha+1}},$$

where $x > 0, \alpha > 0, \theta > 0$, and $\gamma > 0$.
Maximization of the Q-function with respect to α given θ and γ, in the M-step, has a closed form solution given by

$$\hat{\alpha}_k^{(s)} = \frac{n\hat{\pi}_k^{(s)}}{\sum_{i=1}^{n}\pi_{ik}{}^{(s)}\log\left(1+\left(\frac{x_i}{\hat{\theta}_k^{(s)}}\right)^{\hat{\gamma}_k^{(s)}}\right)}.$$

Estimates for θ and γ are obtained based on the marginal weighted log-likelihoods, where $\hat{\alpha}_k$ as a function of $\hat{\theta}_k$ and $\hat{\gamma}_k$ is inserted, and numerical optimization using, for example, the *optim* function from the base package **stats** in R.

Gamma: $X \sim G(\alpha, \theta)$
The density function of the Gamma distribution with shape parameter α and a rate θ is given by

$$f(x|\alpha, \theta) = \frac{\theta^{\alpha}}{\Gamma(\alpha)}x^{(\alpha-1)}e^{-x\theta},$$

where $x > 0, \alpha > 0$, and $\theta > 0$.
Maximization of the Q-function with respect to θ given α, in the M-step, has a closed form solution given by

$$\hat{\theta}_k^{(s)} = \frac{\hat{\alpha}_k^{(s)}\hat{\pi}_k^{(s)}n}{\sum_{i=1}^{n}\pi_{ik}{}^{(s)}x_i}.$$

The estimate for α is obtained based on the marginal weighted log-likelihood, where $\hat{\theta}_k$ as a function of $\hat{\alpha}_k$ is inserted, and numerical optimization.

Inverse Burr: $X \sim IBurr(\tau, \theta, \gamma)$
The density function of the Inverse Burr distribution with shape parameters, τ and γ, and a scale parameter θ, is given by

$$f(x|\tau, \theta, \gamma) = \frac{\tau\gamma\left(\frac{x}{\theta}\right)^{\tau\gamma}}{x\left(1+\left(\frac{x}{\theta}\right)^{\gamma}\right)^{\tau+1}},$$

where $x > 0, \tau > 0, \theta > 0$ and $\gamma > 0$.
Maximization of the Q-function with respect to τ given θ and γ, in the M-step, has a closed form solution given by

$$\hat{\tau}_k^{(s)} = \frac{n\hat{\pi}_k^{(s)}}{\sum_{i=1}^{n}\pi_{ik}{}^{(s)}\log\left(1+\left(\frac{\hat{\theta}_k^{(s)}}{x_i}\right)^{\hat{\gamma}_k^{(s)}}\right)}.$$

The estimates for θ and γ are obtained based on the marginal weighted log-likelihoods, where $\hat{\tau}_k$ as a function of $\hat{\theta}_k$ and $\hat{\gamma}_k$ is inserted, and numerical optimization.

Inverse Gaussian: $X \sim IG(\mu, \lambda)$
The density function of the Inverse Gaussian distribution is given by

$$f(x|\mu, \sigma) = \sqrt{\frac{\lambda}{2\pi x^3}}e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}},$$

where $x > 0, \mu > 0$, and $\lambda > 0$.
Maximization of the Q-function with respect to μ and λ given μ, in the M-step, has a closed form solution given

by

$$\hat{\mu}_k^{(s)} = \frac{\sum_{i=1}^{n} \pi_{ik}{}^{(s)} x_i}{n\hat{\pi}_k^{(s)}}$$

and

$$\hat{\lambda}_k^{(s)} = \frac{n\hat{\pi}_k^{(s)}}{\sum_{i=1}^{n} \pi_{ik}{}^{(s)} \left(\frac{1}{x_i} - \frac{1}{\hat{\mu}_k^{(s)}}\right)}.$$

Log-normal: $X \sim LN(\mu, \sigma^2)$

The density function of the Log-normal distribution, with location parameter, μ, and scale parameter, σ, is given by

$$f(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{(\log(x)-\mu)^2}{2\sigma^2}},$$

where $x > 0$ and $\sigma > 0$.

Maximization of the Q-function with respect to μ and σ given μ, in the M-step, has a closed form solution. These are given by

$$\hat{\mu}_k^{(s)} = \frac{\sum_{i=1}^{n} \pi_{ik}{}^{(s)} \log(x_i)}{\sum_{i=1}^{n} \pi_{ik}{}^{(s)}}$$

and

$$(\hat{\sigma}_k^2)^{(s)} = \frac{\sum_{i=1}^{n} \pi_{ik}{}^{(s)} \left(\log(x_i) - \hat{\mu}_k^{(s)}\right)^2}{\sum_{i=1}^{n} \pi_{ik}{}^{(s)}}.$$

Weibull: $X \sim W(\alpha, \theta)$

The density function of the Weibull distribution with shape parameter, α, and scale parameter, θ, is given by

$$f(x|\alpha, \theta) = \left(\frac{\alpha}{\theta}\right)\left(\frac{x}{\theta}\right)^{(\alpha-1)} e^{-\left(\frac{x}{\theta}\right)^{\alpha}},$$

where $x > 0$, $\alpha > 0$, and $\theta > 0$.

Estimates for α and θ are obtained based on the weighted log-likelihoods and numerical optimization.

In each iteration, the quantities computed in the E-step and M-step are updated until the algorithm converges. The EM algorithm has been shown to increase the log-likelihood values in each iteration thus ensuring convergence in the case of bounded log-likelihoods, even though convergence is not necessary to the global optimum. A suitable stopping criterion is thus to use the relative increase in the log-likelihood function and stop the algorithm if this is smaller than some small pre-specified tolerance value.

2.3. Initialization of the EM algorithm

Because the point to which the EM algorithm converges depends on the initial values, but does not necessarily need to be the global optimum, a good choice of initial values or a good initialization procedure trying out different initial values is crucial in order to ensure that the global optimum is detected. Considering that the likelihood function of finite mixtures is usually multimodal and the EM algorithm is a climbing procedure, an efficient initialization method is important when seeking to find a global maximizer or the best local maximizer in case of an unbounded likelihood function. We consider three different initialization strategies for parameter estimation: (1) Euclidean distance-based, (2) k-means, and (3) Random initialization. One of the objectives of this paper is to compare these three methods in finding good solutions by comparing the best solution found with each of the methods and determining which one can most often find the solution with the highest log-likelihood value.

Each of the initialization methods aims at determining a partition of the data for initializing the EM algorithm.

Euclidean distance-based initialization: This initialization method relies on a measure of distance between cluster centers randomly selected from the data and the observations. The cluster centers are obtained by randomly drawing observations from the data; then a partition of the data is derived by assigning each observation to the closest cluster center based on the Euclidean distance. Maitra (2009) proposed a similar stochastic initialization procedure called RndEM for fitting Gaussian mixtures.

k-means: This initialization strategy is based on the results of a partitioning algorithm that aims at obtaining a partition that optimizes the k-means criterion proposed by Forgy (1965) and MacQueen (1967), i.e., the partition that minimizes the squared Euclidean distance between the observations and their associated cluster centers is selected. The k-means algorithm is one of the most popular methods used in cluster analysis.

Random initialization: McLachlan and Peel (2000) proposed random initialization for Gaussian mixture modeling. This initialization strategy is based on random partitioning of the data into K groups by randomly drawing with equal probability a value from $\{1, \ldots, K\}$ to assign to each of the n observations.

2.4. Computational strategies

Two potential issues may arise when fitting the proposed mixtures with the EM algorithm: (1) unbounded likelihoods and (2) spurious solutions. An unbounded likelihood problem is related to the issue when there exists a path in the parameter space along which the likelihood goes to infinity and ML estimation breaks down. If a mixture component contains very few observations and has only a small scale parameter relative to the other components, then it is referred to as a spurious component. These degenerate solutions usually have higher log-likelihood values than those associated with other local maxima, and they may be selected over competing solutions. In order to reduce the risk to end up on a path leading to an unbounded likelihood and to eliminate spurious solutions, any initial partition that contains less than 1% of the data is disregarded and only partitions meeting this size criterion on the components are used to initialize the EM algorithm.

In the empirical application using the Danish Fire losses and the simulation studies, each initialization strategy is run 100 times and the best solution, as indicated by the maximum log-likelihood, is retained. This best solution provides the initial partition for starting the EM algorithm. The EM algorithm is stopped when the relative difference in log-likelihood values is smaller than 10^{-6} or the maximum number of iterations of 1000 is reached.

2.5. Model selection

For modeling goodness-of-fit of the proposed models, we consider the following measures: negative log-likelihood (NLL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

The negative log-likelihood measure is used to compare the models with the same number of parameters. Let $\ell(\theta)$ denote the log-likelihood function for a given model, then NLL is defined by

$$NLL = -\ell(\theta).$$

In the case of mixture modeling, if the number of components increases, the value of the log-likelihood will improve so that the log-likelihood values as a function of K will be monotonically increasing and thus NLL is only a suitable measure to compare models of the same complexity, i.e., with the same number of parameters.

Comparing models with a different number of parameters requires criteria that penalize the log-likelihood values by adding a term that increases with the number of parameters and that aims at balancing model fit with model complexity.

The Akaike Information Criterion (AIC) is a popular model selection procedure proposed by Akaike (1974). The AIC considers twice the negative log-likelihood plus a penalty term that is equal to twice the number of free parameters (p) in the model, i.e.,

$$AIC = -2\ell(\theta) + 2p.$$

The model with the minimum AIC value is selected as the preferred model to fit the data.

The Bayesian Information Criterion (BIC), proposed by Schwarz (1978), is another commonly used method in model selection. Similar to AIC, the BIC approach adjusts the log-likelihood by a penalty term that considers the number of observations (n) in the sample in addition to the number of parameters in the model, i.e.,

$$BIC = -2\ell(\theta) + p\log(n).$$

The model with the minimum BIC is chosen as the best model to fit the data. The BIC is often preferred in finite mixture modeling (for a discussion, see for example Fraley and Raftery, 2002).

2.6. Risk measures

Following the notation by Klugman et al. (2012), let X denote a random variable and π_p is the $100p$ quantile of the distribution of X. The Value-at-Risk for a random variable X, denoted as $VaR_p(X)$, is the same as π_p and satisfies

$$P(X > \pi_p) = 1 - p. \tag{2.4}$$

In the case of finite mixtures, $VaR_p(X)$ does not have a closed form solution and requires a numerical solution of the following equation

$$F_X(\pi_p) = p, \tag{2.5}$$

where F_X represents the cumulative distribution function of X. This can, for example, be done in R using the function *uniroot()* from the base package **stats**.

For a random variable X, the Tail-Value-at-Risk, denoted as $TVaR_p(X)$, is the conditional expectation of X given that X exceeded the $100p$ quantile of the distribution, i.e.,

$$TVaR_p(X) = E(X|X > \pi_p) = \frac{\int_{\pi_p}^{\infty} xf(x)\,dx}{1 - F_X(\pi_p)} = \frac{\int_{\pi_p}^{\infty} xf(x)\,dx}{1-p}. \tag{2.6}$$

The conditional expectation satisfies the "linearity" property as discussed by Ross (2014). Thus, it follows that the $TVaR_p(X)$ for a mixture will be computed as the weighted sum of the $TVaR_p(X)$ of each of the component distributions. The weights, which correspond to the mixing probabilities, are part of the parameters estimated with the EM algorithm. All formulas for TVaR have closed form solutions.

3. The analysis

3.1. Data

In this paper, we use a well-known data set on Danish Fire losses that has been analyzed by many researchers for more than two decades. The data were collected by Copenhagen Reinsurance and consist of 2492 records over the period 1980–1990. The losses are in millions of Danish Krone and are not adjusted for inflation over time.

Early studies on Danish Fire losses focused on extreme value theory for heavy tailed distributions, e.g., McNeil (1997) and Resnick (1997). Since 2005, there is an increasing trend in developing composite models. At this point, modeling Danish Fire losses using composite models is fairly comprehensive and exhaustive. Coorey and Ananda (2005) introduced a composite Log-normal-Pareto model. Their model is composed of a Log-normal density up to an unknown threshold and a two-parameter Pareto beyond the threshold. This composition is justified because the Pareto model fits well the upper long tail of the distribution that reflects larger losses with smaller frequency while the Log-normal model fits well the smaller losses with higher frequency of occurring. Conditions are imposed on the model parameters to ensure continuity and differentiability at the threshold point.

Scollnik (2007) extended the research on composite Log-normal-Pareto models by recognizing the limitations of the model by Coorey and Ananda (2005) due to its fixed mixing weights. It was argued that this model can be interpreted as a two component mixture model with fixed and a priori known weights, thus making this model rather restrictive and less attractive in practice. The alternative two composite models proposed by Scollnik (2007) included Log-normal-Pareto with mixing weights that were not fixed a priori and the Log-normal-Pareto (Type II) model. Further extensions in this stream of research on composite models were proposed by Pigeon and Denuit (2011), who introduced the threshold value as a random variable in existing Log-normal-Pareto composite models, Nadarajah and Bakar (2014), who suggested a composite Log-normal-Burr model, and Scollnik and Sun (2012), who introduced several composite Weibull–Pareto models for modeling loss severity and other forms of actuarial data.

The Danish Fire losses data set, *danish*, was obtained from the **SMPracticals** package (Davison, 2013) in R. This data set shares many common characteristics with loss insurance data in general such as a heavy right tail and skewness. The basic summary statistics include: 0.3134 (minimum), 1.1572 (first quartile), 3.0630 (mean), 1.6339 (median), 2.6455 (third quartile), and 263.2504 (maximum). We can observe that the mean is much higher than the median as well as the third quartile, indicating extreme skewness in the right tail of the distribution. This is also confirmed by the high value of the skewness coefficient, which equals 19.88. Fig. 1 shows the histogram of the data with the top five extreme losses in the right tail indicated by the arrows.

Actuaries are seeking to fit the best model to this loss data. The best models are praised for their good fit not only in the body, but also in the tail of the distribution. They are selected based on goodness-of-fit criteria such as AIC, BIC, or the log-likelihood. Once the parameters of the best fitting model are determined, these models are employed in a wide range of actuarial applications. Pricing, evaluating risk measures (e.g., value-at-risk or conditional tail expectation), and assessing optimal reinsurance retention levels represent just a few areas of applications. We consider modeling the Danish Fire losses data set using the six mixture models previously introduced.

3.2. Results

In this section we present our results in comparison to the results of the best composite models previously published in the literature as well as a composite Weibull–Burr mixture, i.e., a mixture model consisting of two components where one component follows a Weibull distribution and the other component a Burr distribution. Estimation of the mixing probabilities and component-specific parameters of the composite Weibull–Burr mixture model

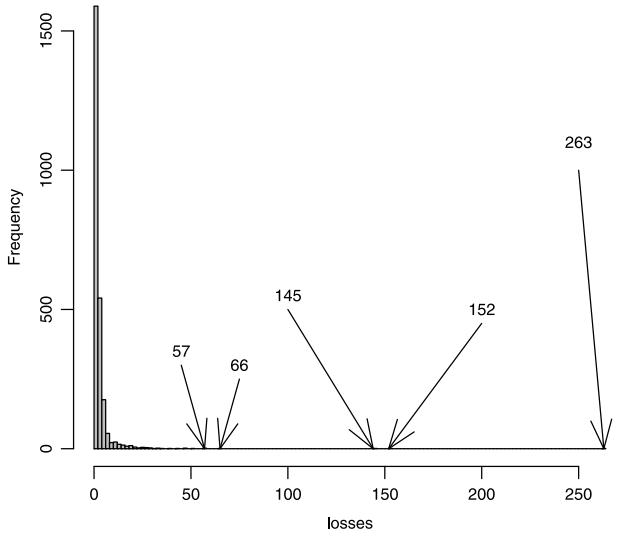

Fig. 1. Danish Fire losses.

is also possible based on the EM algorithm. Given that the M-step has been derived for mixtures of Weibull distributions as well as mixtures of Burr distributions, the extension of the EM algorithm to the case of the composite Weibull–Burr mixture model is straightforward: In the E-step, only the suitable component distribution needs to be evaluated, but this step remains unchanged otherwise. In the M-step, the parameters of each component are determined independently for all considered mixture models. In the case of the composite Weibull–Burr mixture, for each component the component-specific M-step of the corresponding distribution needs to be performed. The details on the implementation of the EM algorithm for the composite Weibull–Burr mixture are included in the Appendix.

The top portion of Table 1 shows the NLL, AIC, and BIC results of the three best-fitting Weibull composite models reported by Bakar et al. (2015) for modeling Danish Fire losses. Below the composite models, Table 1 includes the NLL, AIC, and BIC results of the composite Weibull–Burr mixture. The BIC result for the composite Weibull–Burr mixture is lower than the BIC value of the composite Weibull–Burr model, indicating that the Weibull–Burr mixture provides a better fit to the data. The bottom portion of Table 1 shows how the NLL, AIC, and BIC change for the different mixture models with varying number of mixture components K fitted using our proposed methodology. The results are based on the best model detected over the three different initialization strategies. The results for the optimal number of components, based on BIC, are presented in bold for each model. The models are shown for each component distribution up to this best model plus the model with one component more, while all models with 1 to 8 components were fitted. The 2-component mixture with a Burr distribution in the components has the lowest BIC value, followed by the 3-component Inverse Burr mixture with the second lowest BIC value, and the 5-component Log-normal mixture with the third smallest BIC value. The 2-component Burr mixture has lower NLL, AIC, and BIC values when compared to the best three Weibull composite models published by Bakar et al. (2015): Weibull–Burr, Weibull-Loglogistic, Weibull-Inverse Paralogistic. The 2-component Burr mixture also has a lower BIC value than the composite Weibull–Burr mixture. The parameters of the top three mixture models are summarized in Table 2. Densities for the top three models are shown in Fig. 2. The dashed lines correspond to the weighted individual component densities, while the full line is the overall mixture density. For example, the top panel of Fig. 2 shows the 2-component Burr mixture, with individual weighted Burr components depicted with dashed lines and overall mixture depicted by the full line.

Table 3 summarizes the results for the risk measures. The empirical VaR compares well with the theoretical estimates for the Danish Fire losses data. For the best mixture based on the

Table 1
Models used in modeling Danish Fire losses.

Composite model		NLL	AIC	BIC
Weibull–Burr		3817.570	7645.140	7674.244
Weibull-Loglogistic		3821.229	7650.458	7673.741
Weibull-Inverse Paralogistic		3820.935	7649.870	7673.153
Composite mixture		**NLL**	**AIC**	**BIC**
Weibull–Burr		3804.013	7620.025	7654.950
Mixture	K	**NLL**	**AIC**	**BIC**
Burr	1	3835.119	7676.239	7693.701
	2	**3786.900**	**7587.800**	**7628.546**
	3	3781.873	7585.746	7649.776
Gamma	1	5243.027	10 490.054	10 501.696
	2	4162.040	8334.080	8363.184
	3	3936.038	7888.077	7934.644
	4	3830.164	7682.327	7746.357
	5	**3808.027**	**7644.055**	**7725.546**
	6	3799.091	7632.181	7731.136
Inverse Burr	1	3966.830	7939.661	7957.123
	2	3833.802	7681.604	7722.350
	3	**3780.728**	**7583.457**	**7647.486**
	4	3772.805	7575.611	7662.924
Inverse Gaussian	1	4516.307	9036.614	9048.256
	2	3965.949	7941.897	7971.001
	3	3876.794	7769.588	7816.154
	4	3798.037	7618.074	7682.103
	5	**3784.355**	**7596.709**	**7678.201**
	6	3775.126	7584.252	7683.206
Log-normal	1	4433.891	8871.782	8883.423
	2	3955.789	7921.578	7950.682
	3	3856.247	7728.494	7775.061
	4	3793.160	7608.321	7672.350
	5	**3779.101**	**7586.202**	**7667.694**
	6	3776.051	7586.102	7685.057
Weibull	1	5270.471	10 544.941	10 556.583
	2	4304.567	8619.133	8648.238
	3	4051.493	8118.986	8165.553
	4	3925.195	7872.391	7936.420
	5	3878.676	7785.352	7866.843
	6	3840.235	7714.471	7813.425
	7	**3822.601**	**7685.202**	**7801.619**
	8	3816.725	7679.450	7813.329

Table 2
Danish Fire losses: Parameter estimates for the best mixture models.

	Component				
	1	2	3	4	5
	Burr				
$\hat{\pi}$	0.397634	0.602366			
$\hat{\alpha}$	0.207175	0.028161			
$\hat{\gamma}$	7.047898	50.277542			
$\hat{\theta}$	1.236993	0.856898			
	Inverse Burr				
$\hat{\pi}$	0.178951	0.251028	0.570021		
$\hat{\tau}$	8.535262	2.472452	2.378904		
$\hat{\gamma}$	11.00894	1.684065	3.375687		
$\hat{\theta}$	0.220443	2.060333	0.018502		
	Log-normal				
$\hat{\pi}$	0.107676	0.123036	0.171483	0.249406	0.348399
$\hat{\mu}$	−0.072684	1.855681	0.117780	1.080790	0.464218
$\hat{\sigma}$	0.064281	1.024611	0.122131	0.427895	0.220077

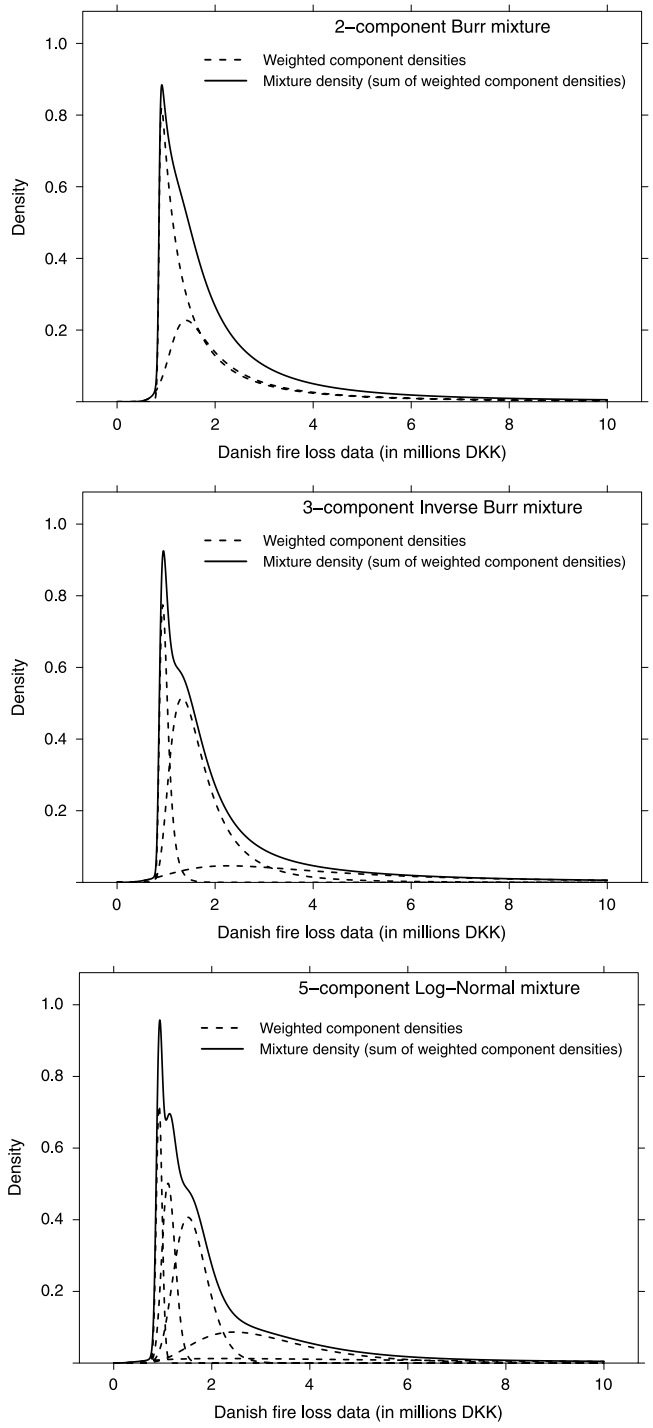

Fig. 2. The three best-fitting mixture models characterized by their mixture densities (full lines) as well as weighted component densities (dashed lines) fitted to the Danish Fire losses data set.

Burr distribution, the relative difference between empirical and theoretical VaR is 1.7%. The second smallest relative difference of 4.2% can be observed for the Inverse Burr mixture. The relative difference reported by Bakar et al. (2015, p. 152) for the composite Weibull–Burr model was 2.3%.

Due to high skewness in the tail of the distribution, the theoretical TVaR is expected to be different from the empirical result. The result for the two component Burr mixture is similar to the result reported by Bakar et al. (2015) for the composite Weibull–Burr model since a Burr component, in both modeling approaches, is used to model the tail of the distribution. The

Table 3
Danish Fire losses: Summary of risk measures.

Empirical estimates	VaR(0.99)	TVaR(0.99)
	24.61	54.60
Composite model	**VaR(0.99)**	**TVaR(0.99)**
Weibull–Burr	25.18	82.59
Weibull-Loglogistic	22.70	62.80
Weibull-Inverse Paralogistic	22.60	60.35
Composite mixture	**VaR(0.99)**	**TVaR(0.99)**
Weibull–Burr	31.14	146.22
Mixture models	**VaR(0.99)**	**TVaR(0.99)**
2-*K* Burr	25.02	82.32
3-*K* Inverse Burr	23.57	58.62
5-*K* Log-normal	26.75	47.22
5-*K* Inverse Gaussian	29.78	51.46
5-*K* Gamma	31.46	45.99
7-*K* Weibull	31.84	47.64

Inverse Burr mixture has the smallest relative difference of 7.4% in TVaR among the top three models that we selected for modeling Danish Fire losses data. In addition, the composite mixture model returns the highest value for the TVaR, which is observed in combination with one of the highest theoretical VaR values and which considerably exceeds the empirical TVaR.

Table 4 summarizes the results of three goodness-of-fit tests for the top three mixture models. These tests are: Kolmogorov–Smirnov, Anderson–Darling, and Chi-Square tests (compare, for example, Lee and Lin, 2010). The test statistic for each test is presented with its corresponding *p*-value given in parentheses. The results of all three goodness-of-fit tests indicate that the fitted distributions are an appropriate representation of the population, as hypothesized under the null hypothesis; in none of the tests the null hypothesis is rejected at a 5% significance level. Fig. 3 shows the QQ- and PP-plots for the top three fitted mixture models, indicating also a very good fit for nearly all observations in the data set. Overall, these diagnostic tools provide additional evidence that the proposed methodology provides a good fit for the Danish fire data if suitable mixture models are selected.

A comparison of the initialization methods was done for Danish Fire losses on the basis of comparing the best log-likelihood values obtained over all repetitions. A basic statistical summary is determined for the relative differences between the optimal log-likelihood value obtained over all three initialization methods and the best log-likelihood value for each initialization method for each mixture model and number of components $K \geq 2$. This comparison indicates that the Euclidean distance-based initialization method performs best, followed by random initialization, and *k*-means. The Euclidean distance-based initialization method obtained the best solution 57% of the time, followed by random initialization with 30%, and *k*-means initialization with 13%. Even if the Euclidean distance-based initialization method did not obtain the best solution among the three initialization methods, only slightly worse results were obtained with the relative maximum deviation detected in comparison to the best log-likelihood value obtained being less than 0.8%. By comparison, these maximum relative deviations are 3.2% for the random initialization method and 3.5% for the *k*-means initialization method.

3.3. Simulation study

In this section we provide the experimental validation of the proposed method based on a simulation study. Because finite mixtures of distributions are known to be a flexible tool for approximating arbitrary distribution functions in a semi-parametric way,

Table 4
Results of goodness-of-fit tests with their test statistics (corresponding p-values in parentheses).

Mixture model	Kolmogorov–Smirnov	Anderson–Darling	Chi-square	Accepted at $\alpha = 0.05$
2-*K* Burr	0.01591 (0.5537)	0.50013 (0.7467)	22.446 (0.2626)	Yes
3-*K* Inverse Burr	0.00969 (0.9734)	0.27093 (0.9583)	22.639 (0.2536)	Yes
5-*K* Log-normal	0.00874 (0.9912)	0.16063 (0.2029)	23.827 (0.9977)	Yes

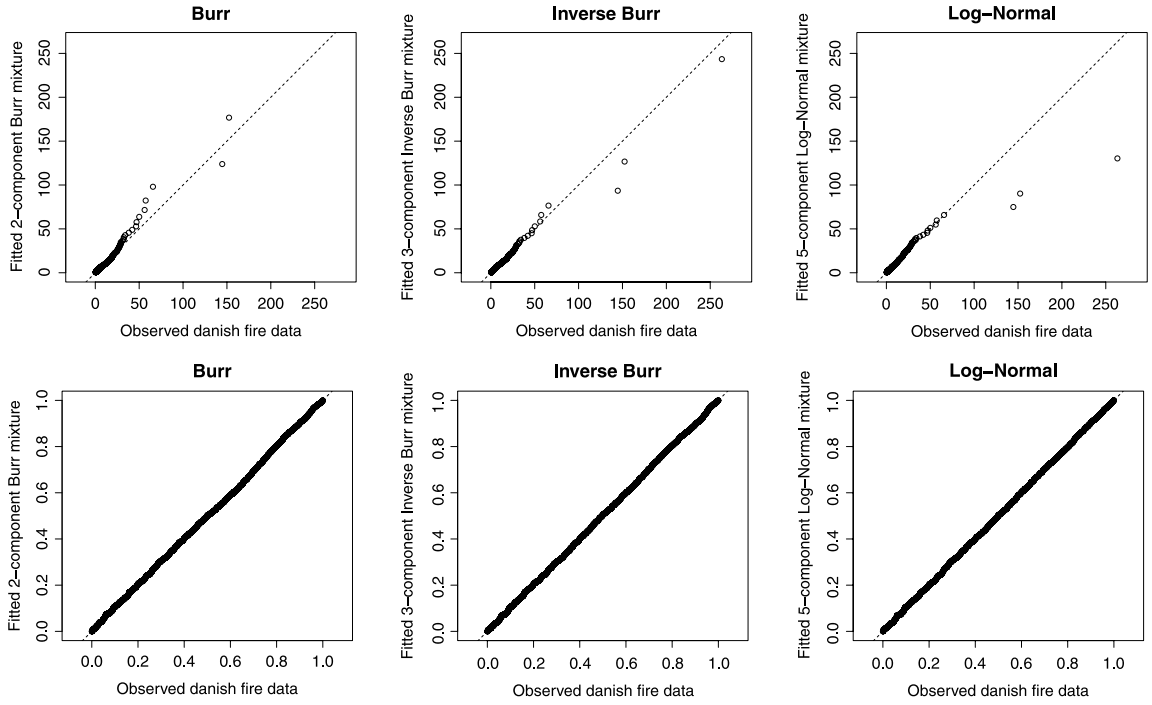

Fig. 3. QQ-plots (top row) and PP-plots (bottom row) for the best three mixture models.

we focus in the simulation study on investigating how well the different mixture distributions are able to approximate data following a distribution likely to be encountered in insurance applications and how close these mixture distributions are for this kind of data.

The simulation study is designed to evaluate the performance of fitting the six proposed mixture models to data from one of the mixture of distributions considered in this paper. This means data is sampled from six different mixture models, each having a different component distribution and constituting a different simulation setting. We simulated 50 samples from 2-component mixtures from each of the parametric families investigated in the paper. The size of each sample corresponds to the size of Danish Fire loss data (2492 observations). The parameters of the 2-component models are based on the best-fitting solutions obtained when modeling the Danish Fire loss data. The mixture models drawn from thus aim at mimicking the same data distribution induced by the Danish Fire losses data and can be assumed to be rather close such that they can also be approximated well by other mixture models. All six mixture models were fitted with the EM algorithm used in combination with the three initialization strategies to each data set with the number of components varying from 1 to 6.

The BIC value is determined for each model fitted to the different data sets and with the different parametric distributions for the components. In addition, the BIC value is determined for each data set and parametric distribution based on the true mixture model, i.e., the mixture model with the parameter values drawn from. This allows us to assess (1) if the estimation based on the EM algorithm with the different initialization methods is able to obtain similar or even better BIC values, which would indicate that the suitable optimum was detected and (2) if a given mixture model with a certain parametric component distribution can also be approximated well by other mixture models and which mixture models they are. Based on these BIC values, we compute the relative difference, ${RD_{ij'j}}^{BIC}$, in BIC between the best fitting model with the distribution for the components fixed and the "true" model as follows

$$ {RD_{ij'j}}^{BIC} = \frac{[\min_{kl}({BIC_{ijkl}}^{fit}) - {BIC_{ij'}}^{true}]}{{BIC_{ij'}}^{true}}, $$

where $i = 1, \ldots, 50$ represents the number of samples, $j' = 1, \ldots, 6$ denotes the six different distributions of the models where the data is drawn from, $j = 1, \ldots, 6$ denotes the six different distributions of the models fitted, $k = 1, \ldots, 6$ denotes the number of components of the fitted models, and $l = 1, 2, 3$ denotes the initialization strategy index.

Further, the relative differences, ${RD_i}^{VaR}$ and ${RD_i}^{TVaR}$, between the fitted 2-component mixture model with the correct component distribution and the true model on the basis of VaR and TVaR, are computed as follows

$$ {RD_{ij}}^{VaR} = \frac{{VaR_{ijl^*(i,j)}}^{fit} - {VaR_{ij}}^{true}}{{VaR_{ij}}^{true}}, $$

$$ {RD_{ij}}^{TVaR} = \frac{{TVaR_{ijl^*(i,j)}}^{fit} - {TVaR_{ij}}^{true}}{{TVaR_{ij}}^{true}}, $$

Table 5
Simulation results for true fitted versus true model.

	BIC		VaR		TVaR	
	Mean	SD	Mean	SD	Mean	SD
Log-normal	−0.0007	0.0004	0.0134	0.0532	0.0148	0.0642
Inverse Gaussian	−0.0006	0.0004	0.0097	0.0559	0.0080	0.0610
Gamma	−0.0006	0.0004	0.0008	0.0480	0.0002	0.0500
Weibull	0.0007	0.0031	0.0002	0.0491	−0.0008	0.0541
Burr	−0.0008	0.0004	0.0293	0.0938	0.1209	0.2485
Inverse Burr	−0.0027	0.0008	0.0028	0.0824	0.0111	0.1378

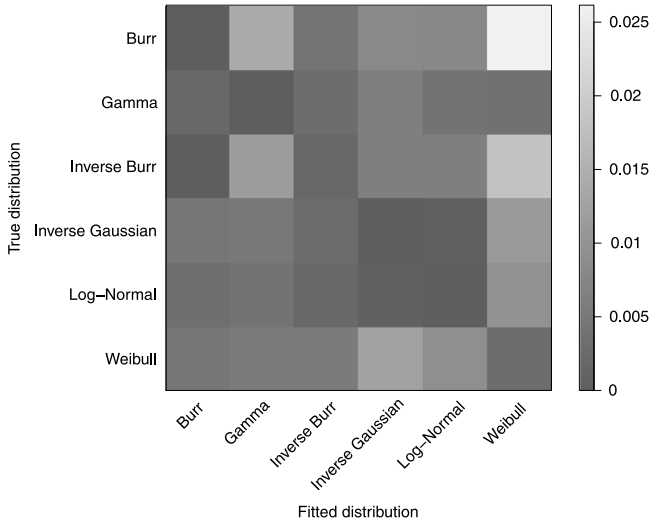

Fig. 4. Relative difference in BIC between true model and best fitting model for each of the different parametric distributions in the components.

where $i = 1, \ldots, 50$ denotes the number of data sets, $j = 1, \ldots, 6$ the six parametric distributions and $l^*(i, j)$ denotes the initialization method giving the minimum BIC for data set i and distribution j.

The mosaic plot in Fig. 4 shows how close the best fitting distributions for each of the parametric distributions are to the true model according to the BIC values. Darker color indicates smaller $RD_{ij'j}^{BIC}$ values, i.e., a stronger concordance between the best fitting distribution and the true model. The mixtures with parametric distributions Burr, Gamma, and Weibull are best fitted with models using the same parametric distribution in the components. When Inverse Burr is the component distribution of the true model, the best fitting distribution is Burr, followed by Inverse Burr. Similarly, when Inverse Gaussian is the component distribution of the true model, Inverse Gaussian and Log-normal are the best fitting models. When Log-normal is the true model, Log-normal and Inverse Gaussian are the best fitting distributions. However, in the last three cases, the relative difference between the best solutions and the solutions obtained for the model with the correct component distribution and number of components is only within 4.7%. These results indicate the flexibility of the mixture approach in also approximating mixtures of other distributions. However, the true model is only slightly outperformed and would thus be nearly always included when, for example, the best three models would be considered, an approach we pursue in the case of modeling and interpreting the results of the Danish Fire losses.

Table 5 provides the summary (mean and standard deviation) of the relative differences in BIC, VaR, and TVaR between the fitted model with 2 components and the correct component distribution and the true model the data is drawn from for the 50 data sets and each distribution. The results for BIC indicate that in general better results are obtained by fitting the model than if the true parameter values are used to evaluate the log-likelihood. This implies that the problem of having only local optima returned by the EM algorithm seems to not be an issue for the initialization methods employed. For VaR and TVaR the distributions of the relative difference between the true fitted and the true model are essentially all unbiased with a very small variability, with the strongest deviations being observable for Burr distribution.

The comparison of the performance of the initialization methods for the results obtained in the simulation study indicates that again Euclidean-based initialization performs best in 41% of the cases, followed by random initialization in 37% of the cases and by k-means initialization in 22% of the cases.

4. Conclusion

In this paper, we proposed modeling insurance losses using K-component finite mixture models. Estimation with the EM algorithm is described and three initialization strategies for the EM algorithm are compared: Euclidean distance-based, k-means, and Random initialization. Six finite mixture models were developed with component-specific distributions from parametric, non-Gaussian families: Burr, Gamma, Inverse Burr, Inverse Gaussian, Log-normal, and Weibull. Risk measures (VaR and TVaR) were calculated for each mixture model. To our knowledge, this approach is the first one to consider mixture modeling outside of the non-Gaussian family of distributions for loss data without imposing any restriction on the parameter estimates, while considering different initialization strategies when fitting the models using the EM algorithm. The methodology developed in this paper is implemented using the statistical computing environment R and is available in the R package **flexmix** (Grün and Leisch, 2008).

Our results were compared to those of composite Weibull models reported by Bakar et al. (2015) for modeling the Danish Fire losses data set. We found that in the case of the Danish Fire losses, 2-component Burr, followed by 3-component Inverse Burr, and 5-component Log-normal mixtures are fitting better than the top composite Weibull models reported by Bakar et al. (2015). We also showed that the computation of risk measures based on the mixture models is straightforward and provides reasonable results compared to composite models. The heterogeneous nature of the insurance claims makes mixture modeling an attractive and flexible approach over composite models for modeling insurance loss data. Our results imply that modeling insurance losses using the K-component mixture of distributions can be effectively employed as a new tool in the area of predictive modeling and risk evaluation. A natural extension of this approach would be to allow also for other parametric distributions in the components.

The flexibility of the mixture modeling approach to approximate general distribution functions has been shown. However, this capability of mixtures is also a drawback when aiming at selecting a single best fitting model. Thus, model uncertainty is an issue when applying this approach and needs to be suitably dealt with. In our empirical analysis, we addressed this problem by considering a set of best fitting models, and we were thus able to obtain robust results by taking the results of all these similarly well-fitting models

into account. An alternative approach to account for model uncertainty and to deduce robust results would be to incorporate the Bayesian model averaging (Hoeting et al., 1999) approach into the mixture modeling framework.

Acknowledgments

The authors thank the Editor, two anonymous Reviewers, and Ann Updike whose helpful suggestions and comments greatly improved the quality of this paper.

Appendix

A.1. The EM algorithm for the composite Weibull–Burr mixture

Consider a random variable $X = (1-\delta)X_1 + \delta X_2$, where $X_1 \sim W(\alpha_1, \theta_1)$, $X_2 \sim Burr(\alpha_2, \theta_2, \gamma)$, $\delta \in \{0, 1\}$, and $P(\delta = 1) = \pi$. The Weibull–Burr mixture is defined as

$$f(x|\boldsymbol{\Psi}) = (1-\pi) f^W(x|\alpha_1, \theta_1) + \pi f^B(x|\alpha_2, \theta_2, \gamma) \tag{A.1}$$

with the parameter vector $\boldsymbol{\Psi} = (\pi, \alpha_1, \theta_1, \alpha_2, \theta_2, \gamma)$ and $0 < \pi < 1$. Distribution functions for Weibull and Burr are denoted as $f^W(x|\alpha_1, \theta_1)$ and $f^B(x|\alpha_2, \theta_2, \gamma)$, respectively. Consider the unobserved latent variable z_i for observation x_i, where

$$z_i = \begin{cases} 0 & \text{if } x_i \text{ is from the Weibull component,} \\ 1 & \text{if } x_i \text{ is from the Burr component.} \end{cases}$$

The complete data log-likelihood function where the observations x_i are augmented with the latent variables z_i is obtained as

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \big[(1-z_i)\log(\pi) + (1-z_i)\log(f^W(x_i|\alpha_1, \theta_1)) + z_i \log(1-\pi) + z_i \log(f^B(x_i|\alpha_2, \theta_2, \gamma))\big].$$

At the sth iteration of the EM algorithm, taking the expectation of $\ell_c(\boldsymbol{\Psi})$ conditional on the observed data and the estimates from the $(s-1)$th iteration results in

$$\begin{aligned} Q(\boldsymbol{\Psi}|\boldsymbol{x}; \boldsymbol{\Psi}^{(s-1)}) &= \mathbb{E}[\ell_c(\boldsymbol{\Psi})|\boldsymbol{x}; \boldsymbol{\Psi}^{(s-1)}] \\ &= \sum_{i=1}^{n} (1-\pi_i^{(s)}) \log(\pi) + (1-\pi_i^{(s)}) \\ &\quad \times \log(f^W(x_i|\alpha_1^{(s-1)}, \theta_1^{(s-1)})) \\ &\quad + \pi_i^{(s)} \log(1-\pi) + \pi_i^{(s)} \\ &\quad \times \log(f^B(x_i|\alpha_2^{(s-1)}, \theta_2^{(s-1)}, \gamma^{(s-1)})), \end{aligned}$$

where $\pi_i^{(s)}$ is the posterior probability that x_i comes from the Burr mixture component given the parameters estimates from iteration $s-1$ and is given by

$$\begin{aligned} \pi_i^{(s)} &= \mathbb{E}[z_i|x_i, \boldsymbol{\Psi}^{(s-1)}] \\ &= \frac{\pi^{(s-1)} f^B\big(x_i|\alpha_2^{(s-1)}, \theta_2^{(s-1)}, \gamma^{(s-1)}\big)}{(1-\pi^{(s-1)}) f^W\big(x_i|\alpha_1^{(s-1)}, \theta_1^{(s-1)}\big) + \pi^{(s-1)} f^B\big(x_i|\alpha_2^{(s-1)}, \theta_2^{(s-1)}, \gamma^{(s-1)}\big)}. \end{aligned}$$

The posterior probability that x_i comes from the Weibull mixture component is given by $1-\pi_i^{(s)}$ in the E-step of the sth iteration.

The M-step consists of finding new estimates for $\boldsymbol{\Psi}$ by maximizing the Q-function. This maximization problem is solved for π in closed form solution: at the sth iteration, the estimate is given by $\hat{\pi}^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \pi_i^{(s)}$. The estimates of the component-specific parameters in $\boldsymbol{\Psi}$ are obtained at the sth iteration based on weighted maximum likelihood estimation using numerical optimization similar to mixtures of Weibull distributions and mixtures of Burr distributions as discussed in detail in Section 2.2.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19 (6), 716–723.
Bakar, S.A.A., Hamzaha, N.A., Maghsoudia, M., Nadarajah, S., 2015. Modeling loss data using composite models. Insurance Math. Econom. 61, 146–154.
Coorey, K., Ananda, M.M., 2005. Modeling actuarial data with a composite lognormal-Pareto model. Scand. Actuar. J. 5, 321–334.
Davison, A., 2013. SMPracticals: Practicals for Use with Davison (2003) "Statistical Models". R package version 1.4–2. URL http://CRAN.R-project.org/package=SMPracticals.
Dayton, C.M., Macready, G.B., 1988. Concomitant-variable latent-class models. J. Amer. Statist. Assoc. 83 (401), 173–178.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM-algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol 39, 1–38.
Forgy, E.W., 1965. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. Biometrics 21, 768–769.
Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis and density estimation. J. Amer. Statist. Assoc. 97 (458), 611–631.
Grün, B., Leisch, F., 2008. FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Softw. 28 (4), 1–35. URL http://www.jstatsoft.org/v28/i04/.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian nodel averaging: A tutorial. Statist. Sci. 14 (4), 382–401.
Keatinge, C.L., 1999. Modeling losses with the mixed exponential distribution. Proc. Casualty Actuar. Soc. 86, 654–698.
Klugman, S.A., Panjer, H.H., Willmot, G.E., 2012. Loss Models: From Data to Decisions, fourth ed. John Wiley & Sons, Hobuken, NJ.
Klugman, S., Rioux, J., 2006. Toward a unified approach to fitting loss models. N. Am. Actuar. J. 10 (1), 63–83.
Lee, S.C.K., Lin, X.S., 2010. Modeling and evaluating insurance losses via mixtures of Erlang distributions. N. Am. Actuar. J. 14 (1), 107–130.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. I. pp. 281–297.
Maitra, R., 2009. Initializing partition-optimization algorithms. IEEE/ACM Trans. Comput. Biol. Bioinform. 6, 144–157.
McLachlan, S., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, Hobuken, NJ.
McNeil, A.J., 1997. Estimating the tails of loss severity distributions using extreme value theory. ASTIN Bull. 27 (1), 117–137.
Nadarajah, S., Bakar, S.A.A., 2014. New composite models for the danish fire insurance data. Scand. Actuar. J. 2, 180–187.
Pfaff, B., McNeil, A., 2012. evir: Extreme values in R. R package version 1.7–3. URL http://CRAN.R-project.org/package=evir.
Pigeon, M., Denuit, M., 2011. Composite lognormal-Pareto model with random threshold. Scand. Actuar. J. 3, 177–192.
R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Resnick, S.I., 1997. Discussion of the danish data on large fire insurance losses. ASTIN Bull. 27, 139–151.
Ross, S.M., 2014. Introduction to Probability Models, eleventh ed. Academic Press.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.
Scollnik, D.P., 2007. On composite lognormal-Pareto models. Scand. Actuar. J. 1, 20–33.
Scollnik, D.P., Sun, C., 2012. Modeling with Weibull-Pareto models. N. Am. Actuar. J. 16 (2), 260–272.
Tijms, H., 1994. Stochastic Models: An Algorithm Approach. John Wiley, Chichester.
Verbelen, R., Antonio, K., Claeskens, G., 2016. Multivariate mixtures of Erlangs for density estimation under censoring. Lifetime Data Anal. 22 (3), 429–453.
Verbelen, R., Gong, L., Antonio, K., Badescu, A., Lin, S., 2015. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. ASTIN Bull. 45 (3), 729–758.
Wang, J.F., Zhau, H.X., Zhou, M.T., Li, L., 2006. A general model for long-tailed network traffic approximation. J. Supercomput. 38 (2), 155–172.