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