Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm

Roel Verbelen*[1], Lan Gong[2], Katrien Antonio[1,3], Andrei Badescu[2], and Sheldon Lin[2]

[1]LStat, Faculty of Economics and Business, KU Leuven, Belgium.
[2]Department of Statistical Sciences, University of Toronto, Canada.
[3]Faculty of Economics and Business, University of Amsterdam, The Netherlands.

July 8, 2015

Abstract

We discuss how to fit mixtures of Erlangs to censored and truncated data by iteratively using the EM algorithm. Mixtures of Erlangs form a very versatile, yet analytically tractable, class of distributions making them suitable for loss modeling purposes. The effectiveness of the proposed algorithm is demonstrated on simulated data as well as real data sets.

Keywords: Mixture of Erlang distributions with a common scale parameter; Censoring; Truncation; Expectation-maximization algorithm; Maximum likelihood.

1 Introduction

The class of mixtures of Erlang distributions with a common scale parameter is very flexible in terms of the possible shapes of its members. Tijms (1994, p. 163) shows that mixtures of Erlangs are dense in the space of positive distributions in the sense that there always exists a series of mixtures of Erlangs that weakly converges, i.e. converges in distribution, to any positive distribution. As such, any continuous distribution can be approximated by a mixture of Erlang distributions to any accuracy. Furthermore, via direct manipulation of the Laplace transform, a wide variety of distributions whose membership in this class is not immediately obvious can be written as a mixture of Erlangs. The class of mixtures of Erlangs with a common scale is also closed under mixture, convolution and compounding. At the same time, it is possible to work analytically with this class leading to explicit expressions for e.g. the Laplace transform, the hazard rate, a Tail-Value-at-Risk (TVAR) and stop-loss moments. A quantile or a Value-at-Risk (VaR) can be obtained by numerically inverting the cumulative distribution function. Klugman et al. (2013), Willmot and Lin (2011) and Willmot and Woo (2007) give an overview of these analytical and computational properties of mixtures of Erlangs.

*Corresponding author. E-mail adress: roel.verbelen@kuleuven.be

Mixtures of Erlang distributions have received most attention in the field of actuarial science. Modeling data on claim sizes is crucial when pricing insurance products. Actuarial models help insurance companies to assess the risk associated with the portfolio, to set the level of premiums (Frees and Valdez, 2008) and reserves (Antonio and Plat, 2014), to determine optimal reinsurance levels (Beirlant et al., 2004) or to determine capital requirements for solvency purposes (Bolancé et al., 2012). Insurance data are often modeled using a parametric distribution such as a gamma, lognormal or Pareto distribution. The usual way to proceed in loss modeling, pricing and reserving is to calibrate the data using several of these parametric distributions and then select, among these, the most appropriate model based on a model selection tool (Klugman and Rioux, 2006). These classes of distributions may however not always be flexible enough in terms of the possible shapes of their members in order to obtain a satisfying fit (e.g. in the presence of multimodal data) and resulting models become intractable when aggregating risks in an insurance portfolio or arising from multiple lines of losses. Ideally, it would be useful to have a single approach to fitting loss models (Klugman and Rioux, 2006) with on the one hand the flexibility of nonparametric density estimation techniques to describe the insurance losses and on the other hand the feasibility to analytically quantify the risk. This is exactly what the class of mixtures of Erlangs has to offer. In particular, using these distributions in aggregate loss models leads to an analytical form of the corresponding aggregate loss distribution, which avoids the need for simulations to evaluate the model.

Mixture models are often used to reflect the heterogeneity in a population consisting of multiple groups or clusters (McLachlan and Peel, 2001). In some applications, these clusters can be physically identified and used to interpret the fitted distributions. This is however not the approach we follow; the components in the mixture will not be identified with existing groups. Mixtures of Erlangs are discussed here for their great flexibility in modeling data and should be regarded as a semiparametric density estimation technique. The densities in the mixture are parametrically specified as Erlangs, whereas the associated weights form the nonparametric part. The number of Erlangs in the mixture with non-zero weights can be viewed as a smoothing parameter. Mixtures of Erlangs have much of the flexibility of nonparametric approaches and furthermore allow for tractable results.

The expectation-maximization (EM) algorithm, first introduced by Dempster et al. (1977), is an iterative method used to compute maximum likelihood (ML) estimates when the data can be viewed as being incomplete and direct maximization of the incomplete data likelihood is either not desirable or not possible (McLachlan and Krishnan, 2008). The EM algorithm is particularly useful in estimating the parameters of a finite mixture. The clue is to view data from a mixture as being incomplete since the associated component-label vectors are not available (McLachlan and Peel, 2001).

Lee and Lin (2010) iteratively use the EM algorithm (Dempster et al., 1977) for finite mixtures to estimate the parameters of a mixture of Erlang distributions with a common scale parameter. For a specified fixed set of shapes, the E- and M-step can be solved analytically without using any optimization method. This makes the EM algorithm for mixtures of Erlangs a pure iterative algorithm which is therefore simple, effective and easy to implement. The initialization is based on Tijm's proof of the denseness property of mixtures of Erlangs (Tijms, 1994, p. 163) which ensures good starting values and fast convergence. Since the number of Erlangs in the mixture and the corresponding shape parameters are pre-fixed and hence not estimated, Lee and Lin (2010) propose an adjustment procedure to identify the 'optimal' number of Erlang distributions and the 'optimal' shape parameters of these distributions in the mixture. The authors illustrate

the flexibility of mixtures of Erlangs by generating data from parametric models (such as the uniform, lognormal, and generalized Pareto distribution) and by approximating the underlying distribution of this sample using a mixture of Erlangs. They further demonstrate the usefulness of mixtures of Erlangs in the context of quantitative risk management for the insurance business. However, modeling censored and/or truncated losses is not covered by the approach in Lee and Lin (2010).

In many practical problems data are censored and/or truncated, for example, due to the way how the data is collected or measured or by the design of the experiment. Censoring entails that you only know in which interval an observation of a variable lies without knowing the exact value while truncation implies that you only observe values that lie within a given range. Interest however is in the underlying distribution of the uncensored and untruncated data instead of the observed censored and/or truncated data. Hence the censoring and truncation has to be accounted for in the analysis.

Survival analysis is the most common application in which data are often censored and truncated. A typical example is a medical study in which one follows patients over a period of time. In case the event of interest has not yet occurred before the end of the study, the patient drops out of the study or dies from another cause, independent of the cause of interest, the event time is right censored. In case the event of interest is known to have occurred between two dates, but the precise date is not known, the event time is interval censored. In actuarial science, insurance losses are often censored and truncated due to policy modifications such as deductibles (left truncation) and policy limits (right censoring). Left truncation is also present in life insurance where members of pension schemes and holders of insurance contracts only enter a portfolio at a certain adult age. Censored and truncated data occur in the context of claim reserving as well (Antonio and Plat, 2014). Indeed, the reserving actuary wants to predict the future development of claims when setting aside reserves at the present moment and has to deal with claims being reported but not yet settled (RBNS) and claims being incurred but not yet reported (IBNR). In operational risk, data are left truncated as they are only recorded in case they exceed a certain threshold. Badescu et al. (2015) use the EM algorithm to fit the correlated frequencies of such left truncated operational loss data using an Erlang-based multivariate mixed Poisson distribution.

Motivated by the large number of areas where censored and truncated data are encountered, the objective in this paper is to develop an extension of the iterative EM algorithm of Lee and Lin (2010) for fitting mixtures of Erlangs with common scale parameter to censored and truncated data. The traditional way of dealing with (grouped and) truncated data using the EM algorithm involves treating the unknown number of truncated observations as a random variable and including it into the complete data vector (Dempster et al., 1977; McLachlan and Krishnan, 2008, p. 66; McLachlan and Peel, 2001, p. 257; McLachlan and Jones, 1988). We do not follow this approach and rather only include the uncensored observations and the component-label vectors in the complete data vector as is also done in Lee and Scott (2012). The fitting procedure is applicable to a wide range of applications. We demonstrate its use in actuarial science and econometrics. Our `R` implementation and additional examples from other domains such as biostatistics are available online at `www.feb.kuleuven.be/roel.verbelen`.

In the following, we briefly introduce mixtures of Erlangs with a common scale parameter in Section 2. The adjusted EM algorithm, able to deal with censored and truncated data, is presented in Section 3. The procedures used to initialize the parameters, to adjust the shapes of the Erlangs in the mixture and to choose the number of components are discussed in Section

4. Examples follow in Section 5 and Section 6 concludes.

2 Mixtures of Erlangs with a common scale parameter

The Erlang distribution is a positive continuous distribution with density function

$$f(x;r,\theta) = \frac{x^{r-1}e^{-x/\theta}}{\theta^r(r-1)!} \qquad \text{for } x>0\,, \tag{1}$$

where r, a positive integer, is the shape parameter and $\theta > 0$ the scale parameter (the inverse $\lambda = 1/\theta$ is called the rate parameter). The cumulative distribution function is obtained by integrating (1) by parts r times

$$F(x;r,\theta) = \int_0^x \frac{z^{r-1}e^{-z/\theta}}{\theta^r(r-1)!}dz = 1 - \sum_{n=0}^{r-1} e^{-x/\theta}\frac{(x/\theta)^n}{n!}\,. \tag{2}$$

Following Lee and Lin (2010) we consider mixtures of M Erlang distributions with common scale parameter $\theta > 0$ and having density

$$f(x;\boldsymbol{\alpha},\boldsymbol{r},\theta) = \sum_{j=1}^{M}\alpha_j\frac{x^{r_j-1}e^{-x/\theta}}{\theta^{r_j}(r_j-1)!} = \sum_{j=1}^{M}\alpha_j f(x;r_j,\theta) \qquad \text{for } x>0\,, \tag{3}$$

where the positive integers $\boldsymbol{r} = (r_1,\ldots,r_M)$ with $r_1 < \ldots < r_M$ are the shape parameters of the Erlang distributions and $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_M)$ with $\alpha_j > 0$ and $\sum_{j=1}^{M}\alpha_j = 1$ are the weights used in the mixture. Similarly, the cumulative distribution function can be written as a weighted sum of terms (2) or (22).

Tijms (1994, p. 163) shows that the class of mixtures of Erlang distributions with a common scale parameter is dense in the space of distributions on $\mathbb{R}^+$. The formulation of the Theorem is given in Appendix A. Lee and Lin (2010) give an alternative proof using characteristic functions.

3 The EM algorithm for censored and truncated data

Lee and Lin (2010) formulate the EM algorithm customized for fitting mixtures of Erlangs with a common scale parameter to complete data. In an Addendum available online[1], we work out the details of this approach using a notation inspired by McLachlan and Peel (2001) and Lee and Scott (2012), based on zero-one component indicators.

Here, we construct an adjusted EM algorithm which is able to deal with censored and truncated data. We represent a censored sample truncated to the range $[t^l, t^u]$ by $\mathcal{X} = \{(l_i,u_i)|\, i = 1,\ldots,n\}$, where t^l and t^u represent the lower and upper truncation points, l_i and u_i the lower and upper censoring points and $t^l \leqslant l_i \leqslant u_i \leqslant t^u$ for $i = 1,\ldots,n$. $t^l = 0$ and $t^u = \infty$ mean no truncation from below and above, respectively. The censoring status is determined as follows:

[1] See www.feb.kuleuven.be/roel.verbelen

Uncensored:	$t^l \leqslant l_i = u_i =: x_i \leqslant t^u$
Left Censored:	$t^l = l_i < u_i < t^u$
Right Censored:	$t^l < l_i < u_i = t^u$
Interval Censored:	$t^l < l_i < u_i < t^u$

For example, when the truncation interval equals $[t^l, t^u] = [0, 10]$, an uncensored observation at 1 gets denoted by $(l_i, u_i) = (1, 1)$, an observation left censored at 2 by $(l_i, u_i) = (0, 2)$, an observation right censored at 3 by $(l_i, u_i) = (3, 10)$ and an observation censored between 4 and 5 by $(l_i, u_i) = (4, 5)$. Thus, l_i and u_i should be seen as the lower and upper endpoints of the interval that contains observation i.

The parameter vector to be estimated is $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$. The number of Erlangs M in the mixture and the corresponding positive integer shapes $\boldsymbol{r}$ are fixed. The value of M is, in most applications, however unknown and has to be inferred from the available data, along with the shape parameters, see Section 4. The portion of the likelihood containing the unknown parameter vector $\boldsymbol{\Theta}$ is given by

$$\mathcal{L}(\boldsymbol{\Theta}; \mathcal{X}) = \prod_{i \in U} \frac{f(x_i; \boldsymbol{\Theta})}{F(t^u; \boldsymbol{\Theta}) - F(t^l; \boldsymbol{\Theta})} \prod_{i \in C} \frac{F(u_i; \boldsymbol{\Theta}) - F(l_i; \boldsymbol{\Theta})}{F(t^u; \boldsymbol{\Theta}) - F(t^l; \boldsymbol{\Theta})}$$

where U is the subset of observations in $\{1, \ldots, n\}$ which are uncensored and C is the subset of left, right and interval censored observations. In case there is no truncation, i.e. $[t^l, t^u] = [0, \infty]$, the contribution of a left censored observation to the likelihood equals $F(u_i; \boldsymbol{\Theta})$ since $l_i = 0$, of a right censored observation $1 - F(l_i; \boldsymbol{\Theta})$ with $u_i = \infty$, and of an interval censored observation $F(u_i; \boldsymbol{\Theta}) - F(l_i; \boldsymbol{\Theta})$.

The corresponding log likelihood is

$$\begin{aligned} l(\boldsymbol{\Theta}; \mathcal{X}) = &\sum_{i \in U} \ln \left(\sum_{j=1}^{M} \alpha_j f(x_i; r_j, \theta) \right) + \sum_{i \in C} \ln \left(\sum_{j=1}^{M} \alpha_j \left(F(u_i; r_j, \theta) - F(l_i; r_j, \theta) \right) \right) \\ &- n \ln \left(\sum_{j=1}^{M} \alpha_j \left(F(t^u; r_j, \theta) - F(t^l; r_j, \theta) \right) \right), \end{aligned} \tag{4}$$

which is difficult to optimize numerically.

3.1 Truncated mixture of Erlangs

The probability density function evaluated in an uncensored observation x_i after truncation (t^l, t^u) is given by

$$\begin{aligned} f(x_i; t^l, t^u, \boldsymbol{\Theta}) &= \frac{f(x_i; \boldsymbol{\Theta})}{F(t^u; \boldsymbol{\Theta}) - F(t^l; \boldsymbol{\Theta})} \\ &= \sum_{j=1}^{M} \alpha_j \cdot \frac{f(x_i; r_j, \theta)}{F(t^u; \boldsymbol{\Theta}) - F(t^l; \boldsymbol{\Theta})} \\ &= \sum_{j=1}^{M} \alpha_j \cdot \frac{F(t^u; r_j, \theta) - F(t^l; r_j, \theta)}{F(t^u; \boldsymbol{\Theta}) - F(t^l; \boldsymbol{\Theta})} \cdot \frac{f(x_i; r_j, \theta)}{F(t^u; r_j, \theta) - F(t^l; r_j, \theta)} \end{aligned}$$

$$= \sum_{j=1}^{M} \beta_j f(x_i; t^l, t^u, r_j, \theta)\,, \tag{5}$$

for $t^l \leqslant x_i \leqslant t^u$ and zero otherwise. This is again a mixture with mixing weights β_j and component density functions given by, respectively,

$$\beta_j = \alpha_j \cdot \frac{F(t^u; r_j, \theta) - F(t^l; r_j, \theta)}{F(t^u; \boldsymbol{\Theta}) - F(t^l; \boldsymbol{\Theta})} \tag{6}$$

and

$$f(x_i; t^l, t^u, r_j, \theta) = \frac{f(x_i; r_j, \theta)}{F(t^u; r_j, \theta) - F(t^l; r_j, \theta)}\,. \tag{7}$$

The component density functions $f(x_i; t^l, t^u, r_j, \theta)$ are truncated versions of the original component density functions $f(x_i; r_j, \theta)$. The weights β_j are obtained by reweighting the original weights α_j by means of the probabilities of the corresponding component to lie in the truncation interval.

3.2 Construction of the complete data vector

The EM algorithm provides a computationally easy way to fit this finite mixture to the censored and truncated data. The main clue is to regard the censored sample $\mathcal{X}$ as being incomplete since the uncensored observations $\boldsymbol{x} = (x_1, \ldots, x_n)$ and their associated component-indicator vectors $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ with

$$z_{ij} = \begin{cases} 1 & \text{if observation } x_i \text{ comes from } j\text{th component density } f(x_i; t^l, t^u, r_j, \theta) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, M$, are not available. The component-label vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are distributed according to a multinomial distribution consisting of one draw on M categories with probabilities $\beta_1, \ldots, \beta_M$ where

$$P(\boldsymbol{Z}_i = \boldsymbol{z}_i) = \beta_1^{z_{i1}} \ldots, \beta_M^{z_{iM}}$$

for $i = 1, \ldots, n$ with z_{ij} equal to 0 or 1 and $\sum_{j=1}^{M} z_{ij} = 1$. We write

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{\text{i.i.d.}}{\sim} \text{Mult}_M(1, \boldsymbol{\beta})\,.$$

Hence, the latent variables $\boldsymbol{Z}_i$ reveal which component density generated observation x_i. Whereas the unconditional truncated probability density function is given by (5), the conditional truncated probability density function of X_i given $Z_{ij} = 1$ is given by (7).

The complete data vector, $\mathcal{Y} = (x_1, \ldots, x_n, \boldsymbol{z}) = \{(x_i, \boldsymbol{z_i}) | i = 1 \ldots n\}$, contains all uncensored observations x_i and their corresponding mixing component vector $\boldsymbol{z}_i$. The log likelihood of the complete sample $\mathcal{Y}$ then becomes

$$l(\boldsymbol{\Theta}; \mathcal{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \ln \Big(\beta_j f(x_i; t^l, t^u, r_j, \theta) \Big)\,, \tag{9}$$

which has a simpler form than the incomplete log likelihood (4) as it does not contain logarithms of sums. The EM algorithm deals with the censored and truncated data from the mixture of Erlangs with common scale in the following steps.

3.3 Initial step

An initial guess for $\boldsymbol{\Theta}$ is needed to start the algorithm. The closer the starting value is to the true maximum likelihood estimator, the faster the algorithm will converge. Parameter initialization is often the sore point of an EM implementation and the study of good initial estimates is often not feasible and disregarded.

For mixtures of Erlangs however, the denseness property (see Tijms (1994, p. 163) and Appendix A) provides an excellent way of coming up with good initial estimates. In the initial step, we deal with the censoring and truncation in a crude manner. We switch to an initializing dataset, denoted by $\boldsymbol{d}$, in which we treat the left and right censored data points as being observed, i.e. we use u_i and l_i, respectively, and we replace the interval censored data points with the midpoint, i.e. we use $(l_i + u_i)/2$. Based on this initial data, we initialize the parameters θ and $\boldsymbol{\alpha}$ as:

$$\theta^{(0)} = \frac{\max(\boldsymbol{d})}{r_M} \quad \text{and} \quad \alpha_j^{(0)} = \frac{\sum_{i=1}^{n} I\left(r_{j-1}\theta^{(0)} < d_i \leqslant r_j\theta^{(0)}\right)}{n}, \quad \text{for } j = 1, \ldots, M, \tag{10}$$

with $r_0 = 0$ for notational convenience. Inspired by Tijms's formulation of the denseness property, the initial scale $\theta^{(0)}$ is chosen such that $\theta^{(0)} r_M$ equals the maximum data point and the initial weights α_j for $j = 1, 2, \ldots, M$ are set to be the relative frequency of data points in the interval $(r_{j-1}\theta^{(0)}, r_j\theta^{(0)}]$. The truncation is only taken into account to transform the initial values for $\boldsymbol{\alpha}$ into the initial values for $\boldsymbol{\beta}$ via (6).

3.4 E-step

In the kth iteration of the E-step, we take the conditional expectation of the complete log likelihood (9) given the incomplete data $\mathcal{X}$ and using the current estimate $\boldsymbol{\Theta}^{(k-1)}$ for $\boldsymbol{\Theta}$ with

$$\begin{aligned} Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) &= E(l(\boldsymbol{\Theta}; \mathcal{Y}) \mid \mathcal{X}; \boldsymbol{\Theta}^{(k-1)}) \\ &= E\left[\sum_{i \in U} \sum_{j=1}^{M} Z_{ij} \ln\Big(\beta_j f(x_i; t^l, t^u, r_j, \theta)\Big) \,\middle|\, \mathcal{X}; \boldsymbol{\Theta}^{(k-1)} \right] \\ &\quad + E\left[\sum_{i \in C} \sum_{j=1}^{M} Z_{ij} \ln\Big(\beta_j f(X_i; t^l, t^u, r_j, \theta)\Big) \,\middle|\, \mathcal{X}; \boldsymbol{\Theta}^{(k-1)} \right] \\ &= Q_u(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) + Q_c(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}), \end{aligned} \tag{11}$$

where $Q_u(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ and $Q_c(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ are the conditional expectations of the uncensored and censored part of the complete log likelihood, respectively.

Uncensored case. The truncation does not complicate the computation of the expectation for the uncensored data as

$$\begin{aligned} Q_u(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) &= E\left[\sum_{i \in U} \sum_{j=1}^{M} Z_{ij} \ln\Big(\beta_j f(x_i; t^l, t^u, r_j, \theta)\Big) \,\middle|\, \mathcal{X}; \boldsymbol{\Theta}^{(k-1)} \right] \\ &= \sum_{i \in U} \sum_{j=1}^{M} E\left[Z_{ij} \middle| \mathcal{X}; \boldsymbol{\Theta}^{(k-1)} \right] \ln\Big(\beta_j f(x_i; t^l, t^u, r_j, \theta)\Big) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in U}\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \ln\Big(\beta_j f(x_i; t^l, t^u, r_j, \theta)\Big) \\
&= \sum_{i\in U}\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \Big[\ln(\beta_j) + (r_j - 1)\ln(x_i) - \frac{x_i}{\theta} - r_j \ln(\theta) \\
&\quad - \ln((r_j-1)!) - \ln\Big(F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\Big)\Big] \,, \qquad (12)
\end{aligned}
$$

with, for $i \in U$ and $j = 1, \ldots, M$,

$$
\begin{aligned}
{}^{u}z_{ij}^{(k)} &= P(Z_{ij} = 1 \mid x_i, t^l, t^u; \boldsymbol{\Theta}^{(k-1)}) \\
&= \frac{\beta_j^{(k-1)} f(x_i; t^l, t^u, r_j, \theta^{(k-1)})}{\sum_{m=1}^{M} \beta_m^{(k-1)} f(x_i; t^l, t^u, r_m, \theta^{(k-1)})} \\
&\overset{(7)}{=} \frac{\beta_j^{(k-1)} f(x_i; r_j, \theta^{(k-1)}) / \left(F(t^u; r_j, \theta^{(k-1)}) - F(t^l; r_j, \theta^{(k-1)})\right)}{\sum_{m=1}^{M} \beta_m^{(k-1)} f(x_i; r_m, \theta^{(k-1)}) / \left(F(t^u; r_m, \theta^{(k-1)}) - F(t^l; r_m, \theta^{(k-1)})\right)} \\
&\overset{(6)}{=} \frac{\alpha_j^{(k-1)} f(x_i; r_j, \theta^{(k-1)})}{\sum_{m=1}^{M} \alpha_m^{(k-1)} f(x_i; r_m, \theta^{(k-1)})} \,, \qquad (13)
\end{aligned}
$$

where we plugged in definitions (6) and (7) of the weights and components of the truncated mixture in the last two equations in order to express this probability in terms of the original mixing weights and mixing components. The E-step for the uncensored part only requires the computation of the posterior probabilities ${}^{u}z_{ij}^{(k)}$ that observation i belongs to the jth component in the mixture, which remains the same in the truncated case and in the untruncated case.

Censored case. Denote by ${}^{c}z_{ij}^{(k)}$ the posterior probability that observation i belongs to the jth component in the mixture for a censored data point. Then

$$
\begin{aligned}
Q_c(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) &= E\left[\sum_{i\in C}\sum_{j=1}^{M} Z_{ij} \ln\Big(\beta_j f(X_i; t^l, t^u, r_j, \theta)\Big) \,\middle|\, \mathcal{X}; \boldsymbol{\Theta}^{(k-1)}\right] \\
&= \sum_{i\in C} E\left[\sum_{j=1}^{M} Z_{ij} \ln\Big(\beta_j f(X_i; t^l, t^u, r_j, \theta)\Big) \,\middle|\, l_i, u_i, t^l, t^u; \boldsymbol{\Theta}^{(k-1)}\right] \\
&= \sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} E\left[\ln\Big(\beta_j f(X_i; t^l, t^u, r_j, \theta)\Big) \,\middle|\, Z_{ij} = 1, l_i, u_i, t^l, t^u; \theta^{(k-1)}\right] \\
&= \sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} \Big[\ln(\beta_j) + (r_j - 1) E\Big(\ln(X_i) \,\Big|\, Z_{ij} = 1, l_i, u_i, t^l, t^u; \theta^{(k-1)}\Big) \\
&\quad - \frac{1}{\theta} E\Big(X_i \,\Big|\, Z_{ij} = 1, l_i, u_i, t^l, t^u; \theta^{(k-1)}\Big) - r_j \ln(\theta) - \ln((r_j-1)!) \\
&\quad - \ln\Big(F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\Big)\Big] \qquad (14)
\end{aligned}
$$

where we used the tower rule in the third equality. Again using Bayes' rule, we can compute these posterior probabilities, for $i \in C$ and $j = 1, \ldots, M$, as

$$
\begin{aligned}
{}^{c}z_{ij}^{(k)} &= P(Z_{ij} = 1 \mid l_i, u_i, t^l, t^u; \boldsymbol{\Theta}^{(k-1)}) \\
&= \frac{\beta_j^{(k-1)} \left(F(u_i; t^l, t^u, r_j, \theta^{(k-1)}) - F(l_i; t^l, t^u, r_j, \theta^{(k-1)})\right)}{\sum_{j=1}^{M} \beta_j^{(k-1)} \left(F(u_i; t^l, t^u, r_j, \theta^{(k-1)}) - F(l_i; t^l, t^u, r_j, \theta^{(k-1)})\right)} \\
&= \frac{\beta_j^{(k-1)} \left(F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})\right) / \left(F(t^u; r_j, \theta^{(k-1)}) - F(t^l; r_j, \theta^{(k-1)})\right)}{\sum_{j=1}^{M} \beta_j^{(k-1)} \left(F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})\right) / \left(F(t^u; r_j, \theta^{(k-1)}) - F(t^l; r_j, \theta^{(k-1)})\right)} \\
&\overset{(6)}{=} \frac{\alpha_j^{(k-1)} \left(F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})\right)}{\sum_{j=1}^{M} \alpha_j^{(k-1)} \left(F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})\right)} .
\end{aligned} \tag{15}
$$

The expression for the posterior probability in the censored case has the same form as in the uncensored case (13), but with the densities replaced by the probabilities in between the upper and lower censoring points. The terms in (14) for $Q_c(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ containing $E(\ln(X_i)|Z_{ij} = 1, l_i, u_i, t^l, t^u; \theta^{(k-1)})$ will not play a role in the EM algorithm as they do not depend on the unknown parameter vector $\boldsymbol{\Theta}$. The E-step requires the computation of the expected value of X_i conditional on the censoring times and the mixing component $\boldsymbol{Z}_i$ for the current value $\boldsymbol{\Theta}^{(k-1)}$ of $\boldsymbol{\Theta}$:

$$
\begin{aligned}
&E\left(X_i \,\middle|\, Z_{ij} = 1, l_i, u_i, t^l, t^u; \theta^{(k-1)}\right) \\
&\quad = \int_{l_i}^{u_i} x \frac{f(x; r_j, \theta^{(k-1)})}{F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})} dx \\
&\quad = \frac{r_j \theta^{(k-1)}}{F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})} \int_{l_i}^{u_i} \frac{x^{r_j} e^{-x/\theta^{(k-1)}}}{\left(\theta^{(k-1)}\right)^{r_j+1} r_j!} dx \\
&\quad = \frac{r_j \theta^{(k-1)} \left(F(u_i; r_j + 1, \theta^{(k-1)}) - F(l_i; r_j + 1, \theta^{(k-1)})\right)}{F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})} ,
\end{aligned}
$$

for $i \in C$ and $j = 1, \ldots, M$, which has a closed-form expression.

3.5 M-step

In the M-step, we maximize the expected value (11) of the complete data log likelihood obtained in the E-step with respect to the parameter vector $\boldsymbol{\Theta}$ over all $(\boldsymbol{\beta}, \theta)$ with $\beta_j > 0$, $\sum_{j=1}^{M} \beta_j = 1$ and $\theta > 0$. The expressions for $Q_u(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ and $Q_c(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ are given in (12) and (14), respectively. The maximization over the mixing weights $\boldsymbol{\beta}$, requires the maximization of

$$
\sum_{i \in U} \sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \ln(\beta_j) + \sum_{i \in C} \sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} \ln(\beta_j) ,
$$

which can be done analogously as in the uncensored case (see Addendum online). We implement the restriction $\sum_{j=1}^{M} \beta_j = 1$ by setting $\beta_M = 1 - \sum_{j=1}^{M-1} \beta_j$. Setting the partial derivatives at $\boldsymbol{\beta}^{(k)}$ equal to zero implies that the optimizer satisfies

$$
\beta_j^{(k)} = \frac{\sum_{i \in U} {}^{u}z_{ij}^{(k)} + \sum_{i \in C} {}^{c}z_{ij}^{(k)}}{\sum_{i \in U} {}^{u}z_{iM}^{(k)} + \sum_{i \in C} {}^{c}z_{iM}^{(k)}} \beta_M^{(k)} \qquad \text{for } j = 1, \ldots, M-1 .
$$

By the sum constraint we have

$$\beta_M^{(k)} = \frac{\sum_{i \in U} {}^u z_{iM}^{(k)} + \sum_{i \in C} {}^c z_{iM}^{(k)}}{n},$$

and the same form also follows for $j = 1, \ldots, M-1$:

$$\beta_j^{(k)} = \frac{\sum_{i \in U} {}^u z_{ij}^{(k)} + \sum_{i \in C} {}^c z_{ij}^{(k)}}{n} \qquad \text{for } j = 1, \ldots, M\,. \tag{16}$$

The new estimate for the prior probability β_j in the truncated mixture is the average of the posterior probabilities of belonging to the jth component in the mixture. The optimizer indeed corresponds to a maximum since the matrix of second order partial derivatives is negative definite matrix with a compound symmetry structure.

In order to maximize $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ with respect to θ, we set the first order partial derivatives equal to zero (see Appendix B). This leads to the following M-step equation for θ:

$$\theta^{(k)} = \frac{\left(\sum_{i \in U} x_i + \sum_{i \in C} E\left(X_i \,|l_i, u_i, t^l, t^u; \theta^{(k-1)}\right)\right)/n - T^{(k)}}{\sum_{j=1}^{M} \beta_j^{(k)} r_j}, \tag{17}$$

with

$$T^{(k)} = \sum_{j=1}^{M} \beta_j^{(k)} \frac{\left(t^l\right)^{r_j} e^{-t^l/\theta} - \left(t^u\right)^{r_j} e^{-t^u/\theta}}{\theta^{r_j - 1}(r_j - 1)!\left(F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\right)}\Bigg|_{\theta = \theta^{(k)}}.$$

As in the uncensored case, the new estimate $\theta^{(k)}$ in (17) for the common scale parameter θ again has the interpretation of the sample mean divided by the average shape parameter in the mixture, but in the formula for the sample mean, we now take the expected value of the censored data points given the censoring times and subtract a correction term $T^{(k)}$ due to the truncation. However, $T^{(k)}$ in (17) depends on $\theta^{(k)}$ and has a complicated form. Therefore, it is not possible to find an analytical solution and we resort to a Newton-type algorithm to solve (17) numerically using the previous value $\theta^{(k-1)}$ as starting value.

The E- and M-steps are iterated until $l(\boldsymbol{\Theta}^{(k)}; \mathcal{X}) - l(\boldsymbol{\Theta}^{(k-1)}; \mathcal{X})$ is sufficiently small. The maximum likelihood estimator of the original mixing weights α_j for $j = 1, \ldots, M$ can be retrieved by inverting expression (6). This is most easily done by first computing

$$\widetilde{\alpha}_j = \frac{\widehat{\beta}_j}{F(t^u; r_j, \widehat{\theta}) - F(t^l; r_j, \widehat{\theta})} \qquad \text{for } j = 1, \ldots, M\,,$$

where $\widehat{\beta}_j$ and $\widehat{\theta}$ denote the values in the final EM step, and then normalizing the weights such that they sum to 1.

4 Choice of the shape parameters and of the number of Erlangs in the mixture

4.1 Initialization

We start by making an initial choice for the number of Erlangs M in the mixture and set the shapes equal to $r_j = j$ for $j = 1, 2, \ldots, M$. Extending Lee and Lin (2010)[2], we introduce a

[2]We acknowledge the help of Simon Lee who suggested this approach in personal communication.

spread factor s by which we multiply the shapes in order to get a wider spread at the initial step, i.e. $r_j = sj$ for $j = 1, 2, \ldots, M$.

The initialization of θ and $\boldsymbol{\alpha}$ is based on the denseness of mixtures of Erlangs (see (Tijms, 1994, p. 163) and Appendix A), as explained in Section 3.3. Each weight α_j gets initialized as the relative frequency of data points in the interval corresponding to the shape parameter r_j. In case this interval does not contain any data points for some j, the initial weight corresponding to the Erlang in the mixture with shape r_j will be zero and consequently the weight α_j will remain zero at each subsequent iteration. This is clear from the updating scheme (16) in the M-step and the expressions (13) and (15) of the posterior probabilities in the E-step. The shapes r_j with initial weight α_j equal to zero are therefore removed from the mixture at the initial step.

Numerical experiments show that the iterative scheme performs well and results in fast convergence using the above choice of initial estimates for θ and $\boldsymbol{\alpha}$.

4.2 Adjusting the shapes

Since the initial shape parameters are pre-fixed and hence not estimated, the fitted mixture might be sub-optimal. Adjustment of the shape parameters is necessary. Ideally, for a given number of Erlangs M, we want to choose optimal values for the shapes. The choice of the shapes for a given M however is an optimization problem over $\mathbb{N}^M$ which is impossible to solve. We have to resort to a practical procedure which explores the parameter space efficiently in order to obtain a satisfying choice for the shapes.

After applying the EM algorithm a first time to obtain the maximum likelihood estimates corresponding to the initial choice of the shape parameters, we perform stepwise variations of the shapes, each time refitting the scale and the weights using the EM algorithm, and compare the log likelihoods of the results. We hereby follow the procedure proposed by Lee and Lin (2010):

1. Run the algorithm starting from the shapes $\{r_1, \ldots, r_{M-1}, r_M + 1\}$ with initial scale θ and weights $\{\beta_1, \ldots, \beta_{M-1}, \beta_M\}$ equal to the final estimates of the previous execution of the EM algorithm. Repeat this step for as long as the log likelihood improves, each time replacing the old set of parameters by the new ones. This procedure is then applied on the $(M-1)$th shape and so forth until all the shapes are treated.

2. Run the algorithm starting from the shapes $\{r_1 - 1, r_2, \ldots, r_M\}$ with initial scale θ and weights $\{\beta_1, \beta_2, \ldots, \beta_M\}$ the final estimates of the previous execution of the EM algorithm. Repeat this step for as long as the log likelihood improves, each time replacing the old set of parameters by the new ones. This procedure is then applied on the 2nd shape and so forth until all the shapes are treated.

3. Repeat the loops described in the previous steps until the log likelihood can no longer be increased.

Using this algorithm we eventually reach a local maximum of the log likelihood, by which we mean that the fit can no longer be improved by either increasing or decreasing any of the r_j.

4.3 Reducing the number of Erlangs

Too many Erlangs in the mixture will result in an issue of overfitting, which is always a problem in statistical modeling. A decision rule such as Akaike's information criterion (AIC, Akaike, 1974) or Schwartz's Bayesian information criterion (BIC, Schwarz, 1978) helps to decide on the value of M. Models with smaller AIC and BIC values are preferred. Any other information criterion (IC) or objective function could be optimized depending on the purpose for which the model is used.

The problem of testing for the number of components is of both theoretical and practical importance and has attracted considerable attention of many studies over the years and still is a major contemporary issue in a mixture modeling context where the underlying population can be conceptualized as being composed of a finite number of subpopulations. Since mixtures of Erlangs are employed here as a semi-parametric density estimation technique and not as model-based clustering, the commonly used criteria of AIC and BIC are adequate for choosing the number of components (McLachlan and Peel, 2001).

We use a backward stepwise search. As mixtures of Erlangs are dense in the space of positive continuous distributions, we start from a close-fitting mixture of M Erlangs resulting from the shape adjustment procedure described in Section 4.2 and compute the value of the IC. We next reduce the number of Erlangs M in the mixture by deleting the mixture component of which the shape r_j has smallest weight β_j, refit the scale and weights using the EM algorithm and readjust the shapes using the same shape adjustment procedure. If the resulting fit with $M-1$ Erlangs attains a lower value of the IC, the new parameter values replace the old ones. We continue reducing the number of Erlangs in the mixture until the value of the IC does no longer decrease by deleting an additional mixture component.

A backward selection has the advantage of providing initial values close to the maximum likelihood estimates of the new set of shapes which greatly reduces the run time (Lee and Lin (2010)). In contrast, by using a forward stepwise procedure it is not clear which additional shape parameter to use and how the parameters from the previous run can be used to provide useful information on parameter initialization.

As a guideline, we recommend to start from an initial choice for the number of Erlangs M and a spread s resulting in a close-fitting or even overfitting of the data.

4.4 Compare the resulting fit using different initializing parameters

Since the log likelihood has multiple local maxima, the value of the initializing parameters M and s can influence the result. Therefore, it is wise to compare the final fits, after the shape adjustment procedure and reduction of the number of Erlangs using an IC, starting from different choices for the initial number of Erlangs M and/or the spread factor s in the initial step. Tuning of such initializing parameters is common in different numerical algorithms and fitting strategies as well (Hastie et al., 2009). Specifically for the case of mixture of Erlangs, many values for the tuning parameters M and s can lead to a satisfying resulting fit, while using a different mixture of Erlangs representation. This is illustrated in the first data example (Section 5.1, Table 1). In order not to limit the flexibility of the fitting procedure, we do not prefix the value of M and s up front and do not propose any stringent rule. The examples in Section 5 show how a small search for these values is often sufficient to obtain satisfactory results. The freedom of doing an even wider search is left as an option to the user.

5 Examples

The usefulness of the proposed fitting procedure is demonstrated using several examples. A first example involves simulated data from a bimodal distrubution which we censor and truncate allowing us to compare the original density and the entire uncensored and untruncated sample to the fitted mixture of Erlangs. The second example illustrates the use of mixtures of Erlangs to represent right-censored unemployment durations. In the third example, we illustrate the use of mixtures of Erlangs in actuarial science in the context of loss modeling. We fit a mixture of Erlang distribution to truncated claim size data and demonstrate how the fitted mixture can be used to analytically price reinsurance contracts. In the final example, we generate data from a generalized Pareto distribution to explore limitations in modeling heavy-tailed distributions. The examples can be replicated using the `R` code and datasets available on our website[3]. Additional examples can be found there as well.

5.1 Simulated censored and truncated bimodal data

We generate a random sample of 5000 observations from the bimodal mixture of gamma distributions with density function given by

$$f_u(x) = 0.4f(x; r = 5, \theta = 0.5) + 0.6f(x; r = 10, \theta = 1)\,. \tag{18}$$

Next we truncate the data by rejecting all observations beneath the 5% sample quantile or above the 95% sample quantile. The remaining 4500 data points are subsequently being right censored by generating 4500 observations from another mixture of gamma distributions with density function

$$f_{rc}(x) = pf(x; r = 5, \theta = 2/3) + (1 - p)f(x; r = 9, \theta = 1.25)\,, \tag{19}$$

with $p = 0.4$. The resulting data set is composed of 2595 uncensored and 1905 right censored data points, and is used to calibrate the Erlang mixture, keeping the lower and upper truncation into account.

Using the automatic search from Section 4.4 we start from $M = 10$ Erlangs in the mixture and let the spread factor s used in the initial step range from 1 to 10. AIC is used to decide upon the number of Erlangs to use in the mixture as explained in Section 4.3. The right censored data points are treated as being observed at the initialization in (10). The different values of the initializing spread all lead to a different final Erlang mixture, which are reported in Table 1. This illustrates the importance of varying the initial spread. Based on the AIC and BIC values (and plots of the fits not shown here), the different models all represent the data quite well.

[3] `www.feb.kuleuven.be/roel.verbelen`. The `R` code contains the procedures discussed in section 4. An illustration is provided for the first example we consider.

Table 1: Demonstration of initialization and fitting procedure on the data generated from (18). Starting point is a mixture of 10 Erlangs. The initial spread factor s ranges from 1 to 10. The superscripts in the last two columns represent the preference order according to that information criterium.

s	$\boldsymbol{r}$	$\boldsymbol{\alpha}$	θ	AIC	BIC
1	3; 12	0.46; 0.54	0.83	13961.09^{5}	13993.15^{1}
2	4; 14; 18	0.44; 0.34; 0.22	0.63	13956.31^{2}	14001.19^{3}
3	6; 15; 23; 31	0.39; 0.12; 0.35; 0.15	0.41	13959.51^{3}	14017.22^{4}
4	5; 15; 21	0.42; 0.20; 0.38	0.51	13955.61^{1}	14000.50^{2}
5	9; 15; 29; 43; 58	0.23; 0.17; 0.14; 0.31; 0.15	0.22	13961.03^{4}	14031.56^{5}
6	8; 14; 29; 43; 59	0.21; 0.20; 0.15; 0.31; 0.13	0.22	13962.63^{6}	14033.16^{6}
7	14; 23; 34; 45; 58; 74; 96	0.20; 0.17; 0.05; 0.07; 0.14; 0.24; 0.13	0.13	13970.25^{10}	14066.42^{10}
8	10; 16; 24; 40; 55; 69; 89	0.12; 0.18; 0.11; 0.10; 0.16; 0.21; 0.12	0.15	13966.94^{8}	14063.11^{8}
9	11; 18; 28; 46; 63; 79; 101	0.11; 0.19; 0.11; 0.10; 0.17; 0.21; 0.11	0.13	13969.23^{9}	14065.41^{9}
10	13; 21; 32; 50; 67; 84; 107	0.14; 0.18; 0.09; 0.10; 0.17; 0.21; 0.11	0.12	13966.63^{7}	14062.81^{7}

The lowest AIC value was reached using spread factor $s = 4$ with a corresponding mixture of 3 Erlangs. The parameter estimates of this final model are given in Table 2.

Table 2: Parameter estimates of the mixture of 3 Erlangs fitted to the censored and truncated data with underlying density (18).

r_j	α_j	θ
5	0.4206869	0.5081993
15	0.2018598	
21	0.3774533	

In order to verify the goodness-of-fit, we might consider analytical tests such as the Kolmogorov-Smirnov test. However, the form of the test statistic and the corresponding distribution is not at all obvious in a censored and truncated setting. For the case of power-law distributions, Clauset et al. (2009) used Kolmogorov-Smirnov tests to evaluate whether the hypothesized distribution adequately describes the tail. Dufour and Maag (1978) modify the form of the test statistic to allow for truncated and censored data. Guilbaud (1988) derive an exact Kolmogorov-Smirnov test for left-truncated and/or right-censored data. In an actuarial context, Chernobai et al. (2014) discuss goodness-of-fit tests for left-truncated loss samples. We mainly focus on graphical goodness-of fit evaluation in this paper.

A graphical comparison of the fitted distribution and the originally generated data can be found in Figure 1. We compare the fitted mixture of Erlangs density to the true density (18) and a histogram of all 5000 generated data points before truncation and censoring in the left plot in Figure 1. The right plot in Figure 1 compares the truncated mixture of Erlangs density to the true truncated density and a histogram of the 4500 data points after truncation and before censoring. The fitted mixture of Erlangs density shows to be a very close approximation of the true density. Varying the spread from 1 to 10 in the initial mixture of 10 Erlangs is sufficient to obtain a satisfactory result, so there is no need to increase the number of Erlangs in the initial step.

In actuarial practice, loss data can sometimes be of multimodal nature due to the fact that the property and casualty losses often come from multiple sources. Clearly, using standard paramet-

ric distributions will result in unsatisfactory approximations as they are incapable of reflecting the multimodal characteristic. Moreover, applying straightforward estimation techniques may lead to non-convergence issues due to the censoring and truncation. On the contrary, convergence is guaranteed in the presented EM algorithm for mixtures of Erlangs and captures the bimodality of the data very flexibly.

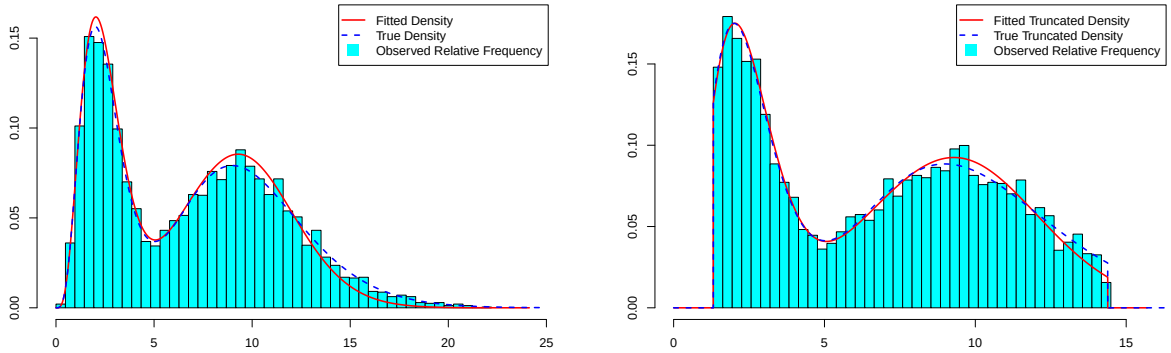

Figure 1: Graphical comparison of the density of the fitted mixture of 3 Erlangs, the true underlying density (18) and the histogram of the generated data before censoring and truncation (left) and of the truncated density of the fitted mixture of 3 Erlangs, the true truncated density and the histogram of the generated data after truncated and before censoring (right).

Next, we investigate the sensitivity with respect to the level of censoring in the data. To that end, we fix the data generated from (18), truncate them at the 5% and 95% sample quantile and vary the value of the mixing weight p in the density (19) of the right censoring distribution from 0 to 1 by 0.1. Let $f(x)$ and $F(x)$ denote the true density and distribution function and $\widehat{f}(x)$ and $\widehat{F}(x)$ the estimated mixture of Erlangs density and distribution function. We measure the performance of both the underlying and the truncated mixture of Erlangs density estimator in approximating the underlying and the truncated true density by calculating the L^1 and L^2 norms:

$$\mathrm{L}^1 = \int_0^\infty \left|\widehat{f}(x) - f(x)\right| \mathrm{d}x \qquad \mathrm{L}_t^1 = \int_{t^l}^{t^u} \left|\frac{\widehat{f}(x)}{\widehat{F}(t^u) - \widehat{F}(t^l)} - \frac{f(x)}{F(t^u) - F(t^l)}\right| \mathrm{d}x$$

$$\mathrm{L}^2 = \left(\int_0^\infty \left(\widehat{f}(x) - f(x)\right)^2 \mathrm{d}x\right)^{1/2} \qquad \mathrm{L}_t^2 = \left(\int_{t^l}^{t^u} \left(\frac{\widehat{f}(x)}{\widehat{F}(t^u) - \widehat{F}(t^l)} - \frac{f(x)}{F(t^u) - F(t^l)}\right)^2 \mathrm{d}x\right)^{1/2}$$

For each value of p in the right censoring distribution (19), we generate 100 censoring samples of size 4500 and each time fit an Erlang mixture to the right censored dataset using the automatic search starting from $M = 10$ Erlangs in the mixture and letting the initial spread s vary from 1 to 10. The averages of the performance measures over the 100 best-fitting resulting mixtures are shown in Table 3. The L^1 and L^2 norms over the truncation interval deteriorate when increasing the censoring level, but remain quite low. This reveals that the performance of the estimator remains excellent when the level of censoring increases, except at the highest level where the estimated Erlang mixture is still bimodal but the second mode and the tail of the

true density are underestimated. The L^1 and L^2 norms over the entire positive real line do not run as parallel with the censoring level as the truncated versions. Note in this context the limitations of accurately estimating the density outside of the truncation interval, since no data has been observed in that region. One should hence not rely on probability statements made using the fitted Erlang mixture outside of the data range.

Table 3: Results of the sensitivity analysis with respect to the level of censoring. For each value of p in the right censoring distribution (19), we generate 100 censoring samples and report the average censoring level and average performance measures of the best-fitting mixtures of Erlang distributions.

p	censoring %	L^1	L^2	L_t^1	L_t^2
0.0	0.2172	0.0862	0.0227	0.0266	0.0097
0.1	0.2695	0.0594	0.0170	0.0280	0.0099
0.2	0.3224	0.0740	0.0197	0.0278	0.0099
0.3	0.3753	0.0864	0.0226	0.0309	0.0109
0.4	0.4289	0.1438	0.0343	0.0329	0.0114
0.5	0.4806	0.1129	0.0277	0.0367	0.0126
0.6	0.5330	0.0905	0.0235	0.0412	0.0140
0.7	0.5844	0.1527	0.0349	0.0465	0.0157
0.8	0.6383	0.1597	0.0377	0.0594	0.0199
0.9	0.6903	0.1787	0.0416	0.0705	0.0236
1.0	0.7426	0.5156	0.1199	0.2276	0.0997

5.2 Unemployment duration

We examine the economic data from the January Current Population Survey's Displaced Workers Supplements (DWS) for the years 1986, 1988, 1990, and 1992 which was first analyzed in McCall (1996). A thorough discussion of this dataset is available in Cameron and Trivedi (2005). The variable under consideration is unemployment duration (`spell`) or more accurately joblessness duration, measured in two-week intervals. All other covariates in the dataset are ignored in the analysis. Following Cameron and Trivedi (2005), a spell is considered complete if the person is re-employed at a full-time job (`CENSOR1` $= 1$) and right-censored otherwise (`CENSOR1` $= 0$). This results in 1073 uncensored data points and 2270 right censored data points.

The parameter estimates of the Erlang mixture, obtained by using the automatic search procedure starting from $M = 10$ Erlangs in the mixture with spread factor s in the initial step ranging from 1 to 10, are given in Table 4. AIC is again used to decide upon the number of Erlangs in the mixture and the right censored data points are treated as being observed at initialization. The lowest AIC value was obtained with a mixture of 8 Erlangs. This optimal choice of shapes was reached using spread factor $s = 10$.

Table 4: Parameter estimates of the mixture of 8 Erlangs fitted to the right-censored unemployment data.

r_j	α_j	θ
8	0.10563305	0.1477264
17	0.09443584	
33	0.08578746	
50	0.09099055	
73	0.04273362	
99	0.14814091	
135	0.07546787	
199	0.35681069	

The Kaplan-Meier estimator (Kaplan and Meier (1958)), also known as the product limit estimator, is the standard non-parametric estimator of the survival function in case of right censored data. The resulting survival curve is a step function with jumps at the observed event times of which the size not only depends on the number of events observed at each event time, but also on the pattern of the censored observations prior to that event time. In order to graphically evaluate the fit, we compare the Kaplan-Meier survival curve, along with 95% confidence bounds, to the survival function of the estimated Erlang mixture in Figure 2. Marks are added on the Kaplan-Meier estimate to indicate censoring times. The fitted survival function provides a smooth fit of the data, closely resembling the non-parametric estimate.

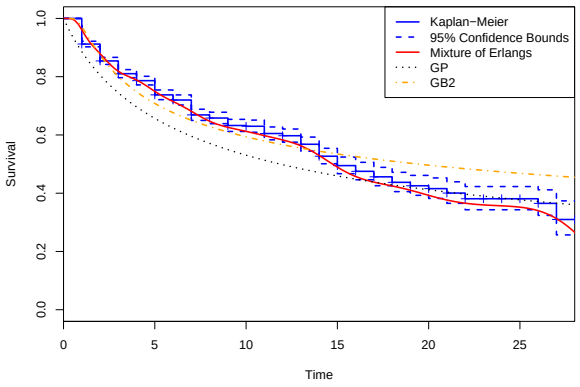

Figure 2: Graphical comparison of the survival function of the fitted mixture of 8 Erlangs and the Kaplan-Meier estimator with 95% confidence bounds for the right-censored unemployment data.

As an illustration, we also compare our approach to two commonly used parametric models, the generalized Pareto distribution (GP) and the generalized beta distribution of the second kind (GB2). In Figure 2, we see how mixtures of Erlangs offer much more flexibility and lead to a more appropriate fit for these data at the cost of requiring more parameters. However, AIC and BIC strongly prefer the mixture of Erlangs approach, see Table 5.

Table 5: Comparison of information criteria for the different models fitted to the right-censored unemployment data.

Model	AIC	BIC
Mixtures of Erlangs	8066.281	8170.230
Generalized Pareto (GP)	8733.718	8745.947
Generalized beta 2 (GB2)	8280.168	8304.627

5.3 Secura Re, Belgian insurance data

The Secura Re dataset discussed in Beirlant et al. (2004) contains 371 automobile claims from 1988 until 2001 gathered from several European insurance companies. The data are uncensored, but left truncated at 1 200 000 since a claim is only reported to the reinsurer if the claim size is at least as large as 1 200 000 euro. The sizes of the claims are corrected among others for inflation. Based on these observations, the reinsurer wants to calibrate a model in order to price reinsurance contracts.

The search procedure using AIC prefers a mixture of only two Erlangs with shapes 5 and 16. The parameter estimates of this best-fitting mixture are shown in Table 6. In Figure 3 (left) we compare the histogram of the truncated data to the fitted truncated density. Figure 3 (right) illustrates that the truncated survival function of the mixture of two Erlangs perfectly coincides with the Kaplan-Meier estimate.

Table 6: Parameter estimates of the mixture of 2 Erlangs fitted to the left-truncated claim sizes in the Secura Re dataset.

r_j	α_j	θ
5	0.97103229	360 096.1
16	0.02896771	

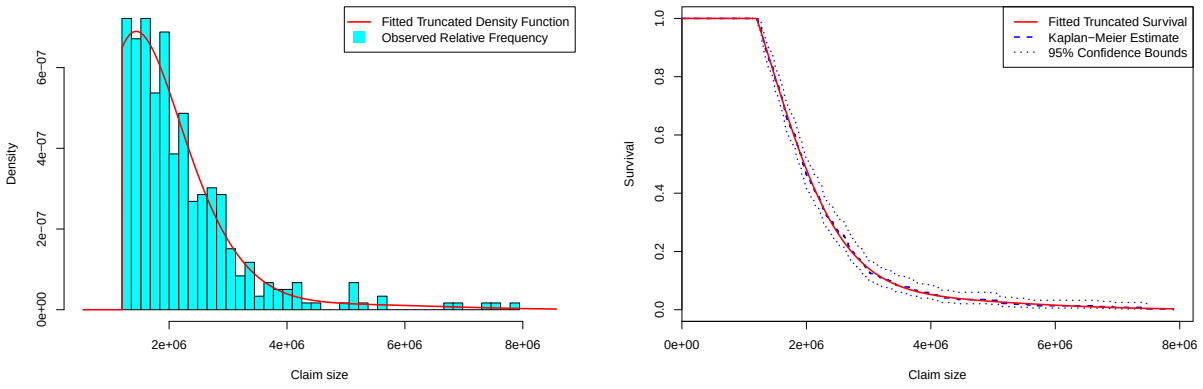

Figure 3: Graphical comparison of the truncated density of the fitted mixture of 2 Erlangs and the histogram of the left-truncated claim sizes (left) and of the truncated survival function and the Kaplan-Meier estimator with 95% confidence bounds (right) for the Secura Re dataset.

In Figure 4, we validate the fit in the tail by plotting the QQ-plot on the left and the log-log plot of the empirical truncated survival function (black dots) and the truncated survival function of the best-fitting Erlang mixture (red line) on the right. Both figures show how the mixture of only two Erlangs achieves a adequate approximation in the tail.

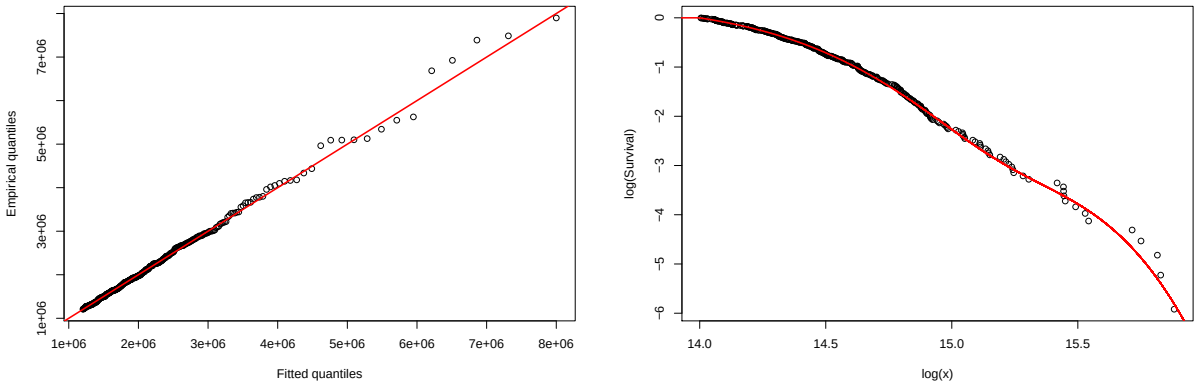

Figure 4: QQ-plot of the empirical quantiles and the quantiles of the fitted mixture of 2 Erlangs with identity line (left) and log-log plot of the empirical truncated survival function and the truncated survival function of the fitted Erlang mixture (right) for the Secura Re dataset.

Following Beirlant et al. (2004, p. 188), we use the calibrated Erlang mixture to price an excess-of-loss (XL) reinsurance contract, where the reinsurer pays for the claim amount in excess of a given limit. The net premium $\Pi(R)$ of such a contract with retention level $R > 1\,200\,000$ is given by

$$\Pi(R) = E((X - R)_+ \mid X > 1\,200\,000)$$

where X denotes the claim size and $(\cdot)_+ = \max(\cdot, 0)$. In case X follows a mixture of M Erlang distributions, where we assume without loss of generality $r_i = i$ for $i = 1, \ldots, M$, the net premium is

$$\begin{aligned}\Pi(R) &= \frac{\theta e^{-R/\theta}}{1 - F(1\,200\,000; \boldsymbol{\alpha}, \boldsymbol{r}, \theta)} \sum_{n=0}^{M-1} \left(\sum_{k=n}^{M-1} A_k \right) \frac{(R/\theta)^n}{n!} \\ &= \frac{\theta^2}{1 - F(1\,200\,000; \boldsymbol{\alpha}, \boldsymbol{r}, \theta)} \sum_{n=1}^{M} \left(\sum_{k=n-1}^{M-1} A_k \right) f(R; n, \theta)\,, \end{aligned} \tag{20}$$

with $A_k = \sum_{j=k+1}^{M} \alpha_j$ for $k = 0, \ldots, M-1$. The derivation of this property can be reconstructed using Willmot and Woo (2007) or Klugman et al. (2013, p. 21). In Table 7, we compare the non-parametric, Hill and Generalized Pareto (GP) based estimates of $\Pi(R)$ for the Secura Re dataset from Table 6.1 in Beirlant et al. (2004, p. 191) to the estimates obtained using formula (20). The maximum claim size observed in the dataset equals 7 898 639 which is the only data point on which the non-parametric estimate of the net premium with retention level $R = 7\,500\,000$ is based. The non-parametric estimate corresponding to retention level $R = 10\,000\,000$ is hence zero. The fitted Erlang mixture allows us to estimate the net premium using intrinsically all data points, but postulates a lighter tail compared to the Pareto-type alternatives since Erlang mixtures have an asymptotically exponential tail (Neuts (1981, p. 62)). Both the estimates

based on the extreme value methodology and those based on the Erlang mixture keep pace with the non-parametric ones, but at the high-end of the sample range, the estimators differ strongly, as implied by the different tail behavior of the three approaches. The reinsurance actuary should carefully investigate the right tail behavior of the data in order to choose his approach.

Table 7: Non-parametric, Hill, GP and Mixture of Erlangs-based estimates for $\Pi(R)$.

R	Non-Parametric	Hill	GP	Mixture of Erlangs
3 000 000	161 728.1	163 367.4	166 619.6	163 987.7
3 500 000	108 837.2	108 227.2	111 610.4	110 118.5
4 000 000	74 696.3	75 581.4	79 219.0	77 747.6
4 500 000	53 312.3	55 065.8	58 714.1	55 746.3
5 000 000	35 888.0	41 481.6	45 001.6	39 451.6
7 500 000	1074.5	13 944.5	16 393.3	4018.6
10 000 000	0.0	6434.0	8087.8	159.6

Besides modeling the tail of the claim size distribution above a certain threshold, Beirlant et al. (2004, p. 198) also estimate a global statistical model to describe the whole range of all possible claim outcomes for the Secura Re dataset. This is needed when trying to estimate $\Pi(R)$ for values of R smaller than the threshold above which the extreme value distribution is fit. Based on the mean excess function, the authors propose the use of a mixture of an exponential and a Pareto distribution (Exp-Par). Instead of having to use this body-tail approach (a form a splicing, see Klugman et al. (2012)) explicitly, the implemented shape adjustment and reduction techniques when fitting the Erlang mixture have guided us to a mixture with two components of which the first one represents the body of the distribution and the second represents the tail. The fitting procedure for Erlang mixtures is able to make this choice implicitly in a data driven way, leading to a close representation of the data. In Table 8 we compare the estimated net premiums from Table 6.2 in Beirlant et al. (2004, p. 198) obtained using the Exp-Par model to the non-parametric and mixture of Erlangs estimates. The estimates based on the fitted Erlang mixture follow the non-parametric ones more closely than those obtained using the Exp-Par model.

Table 8: Non-parametric, Exp-Par and Mixture of Erlangs-based estimates for $\Pi(R)$.

R	Non-Parametric	Exp-Par	Mixture of Erlangs
1 250 000	981 238.0	944 217.8	981 483.1
1 500 000	760 637.6	734 371.6	760 912.9
1 750 000	583 403.6	571 314.1	582 920.1
2 000 000	445 329.8	444 275.5	444 466.6
2 250 000	340 853.2	344 965.2	339 821.4
2 500 000	263 052.7	267 000.7	262 314.6

Note that when $R = 1\,200\,000$, the net premium equals the mean excess loss $E(X - R \mid X > R)$, which is called the mean residual lifetime in survival context. (Klugman et al., 2013, p. 20) show that the distribution of the excess loss or residual lifetime is again a mixture of M Erlangs with

the same scale θ and different weights which we can compute analytically:

$$\alpha_j^* = \frac{\sum_{n=0}^{M-j} \alpha_{n+j} f(R; n+1, \theta)}{\sum_{n=0}^{M-1} A_n f(R; n+1, \theta)} \qquad \text{for } j = 1, \ldots, M\,.$$

5.4 Simulated generalized Pareto data

When modeling claim sizes, the insurer or reinsurer is often confronted with heavy tailed distributions. To safeguard the company against extreme losses that might jeopardize their solvency, an accurate description of the upper tail of the claim size distribution is of utmost importance. In order to explore the limits of Erlang mixtures in approximating heavy-tailed distribution using the presented method, we consider the generalized Pareto distribution with density

$$f_X(x; \mu, \sigma, \xi) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{\left(-\frac{1}{\xi}-1\right)} \qquad \text{for } x \geqslant \mu\,. \tag{21}$$

with location $\mu > 0$, scale $\sigma > 0$ and shape $\xi > 0$. The generalized Pareto family is known for its tail thickness and is used for insurance branches with a high probability of large claims, such as liability insurance. The shape parameter coincides with the extreme value index (EVI) and determines the heaviness of the tail (Beirlant et al., 2004). The higher the value of the EVI, the heavier the tail. The variance is finite for $\xi < 1/2$ and the mean is finite for $\xi < 1$. In general is the kth moment finite for $\xi < 1/k$. When modeling the Secura Re data of the previous example using Pareto-type modeling, Beirlant et al. (2004) estimate the corresponding EVI around 0.3. Using the presented method, we were able to obtain a very good approximation in the tail with a mixture of Erlangs. We now want to illustrate what happens when the EVI further increases, by generating 1000 observations from a generalized Pareto distribution with location $\mu = 10$, scale $\sigma = 2$ and shape $\xi = 1$. In this extreme setting, the EVI equals 1 and none of the moments exist. Location $\mu = 10$ implies that the distribution is left truncated at 10.

In order to obtain a decent approximation of this sample, the initial values of the number of Erlangs M and the spread s become even more important. Due to the fact that the data is very skew and heavy-tailed, the maximum in the dataset is extremely high, i.e. $\max(\boldsymbol{x}) = 10\,636.49$, and many of the initial shape parameters in the mixture will get a corresponding weight equal to zero. To ensure that we start our calibration procedure with sufficient non-zero shape parameters, we decided – after some exploratory choices for M and s – to try all combinations of spread s between 1 and 10 and initial number of Erlangs $M = \left\lceil \frac{\max(\boldsymbol{x})}{i} \right\rceil$ for $i = 1, \ldots, 10$, leading to initial mixtures with 30 to 85 non-zero weight Erlang components. The best-fitting Erlang mixture according to AIC was obtained starting from $M = \left\lceil \frac{\max(\boldsymbol{x})}{7} \right\rceil = 1520$ and $s = 4$, corresponding to a mixture of 34 non-zero weight Erlang components at the initial step. The parameter estimates of the final mixture of 16 Erlangs, after the shape adjustment procedure and the reduction of the number of Erlangs based on AIC, are given in Table 9.

Table 9: Parameter estimates of the mixture of 16 Erlangs fitted to the simulated generalized Pareto data.

r_j	α_j	θ
2	0.9973387302	1.334924
13	0.0016914393	
20	0.0002066144	
28	0.0003513364	
47	0.0001826860	
74	0.0000809294	
120	0.0000458669	
163	0.0000079065	
211	0.0000286491	
286	0.0000073181	
488	0.0000073471	
613	0.0000219147	
3338	0.0000073155	
4472	0.0000073155	
6307	0.0000073155	
7964	0.0000073155	

The underlying untruncated mixture contains 16 components and is dominated by an Erlang distribution with shape 2, modeling the main bulk of the data, whereas the approximation of the tail requires a combination of 15 Erlangs with shapes ranging from 13 to 7964. A graphical comparison of the fitted Erlang mixture and the underlying true distribution up to the 95% sample quantile is shown in Figure 5. The QQ-plot in Figure 6 (left) shows that this mixture does a great job in fitting the sample in the tail. However, the log-log plot of the empirical truncated survival function and the truncated survival function of the best-fitting Erlang mixture in Figure 6 (right) reveals that this approximation is obtained by letting separate Erlang components with a very small weight coincide with the largest data points that lie very far apart. Moreover, all moments of a finite mixture of Erlangs are finite whereas the expected value of the underlying distribution is infinite. We thus conclude that in this extreme setting with EVI equal to 1, the fitted finite mixture of Erlang distributions follows the observed dataset closely, but is not able to extrapolate the heaviness in the tail in comparison to the extreme value methodology based on the Fisher-Tippett-Gnedenko theorem.

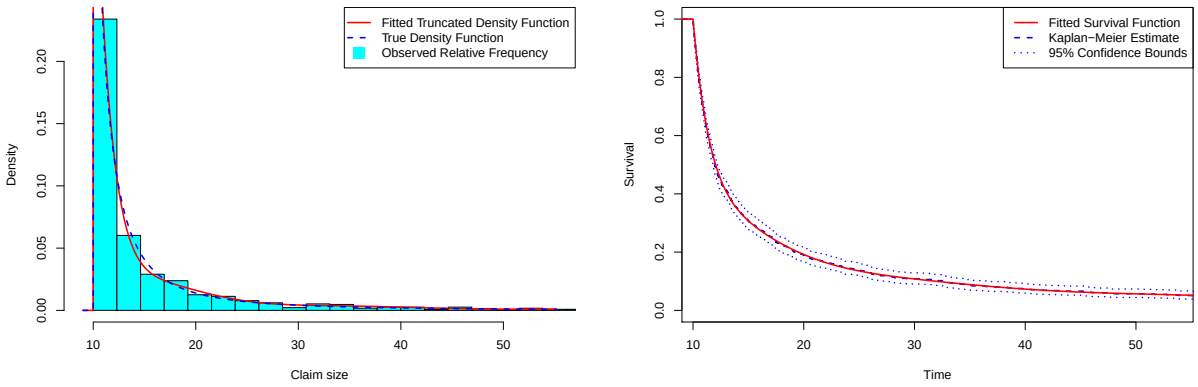

Figure 5: Graphical comparison of the truncated density of the fitted mixture of 16 Erlangs and the histogram (left) and of the truncated survival function and the Kaplan-Meier estimator with 95% confidence bounds (right) for the simulated generalized Pareto data up to the 95% empirical quantile.

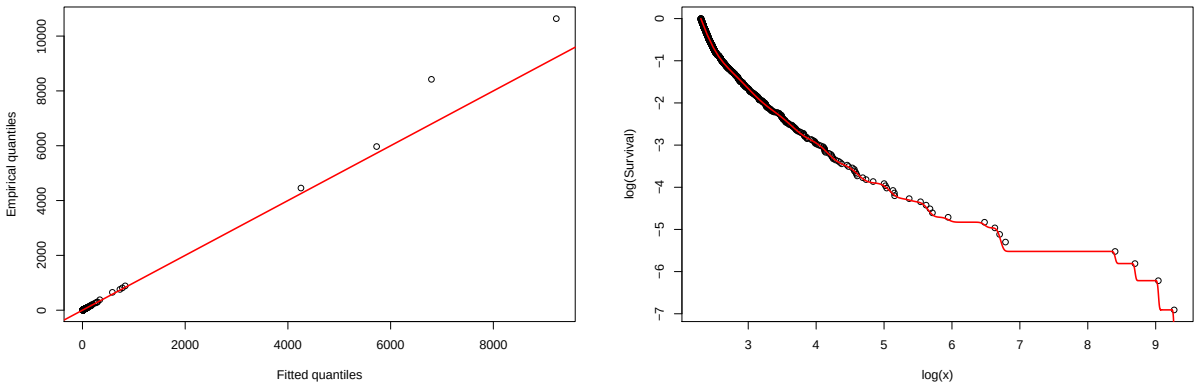

Figure 6: QQ-plot of the empirical quantiles and the quantiles of the fitted mixture of 16 Erlangs with identity line (left) and log-log plot of the empirical truncated survival function and the truncated survival function of the fitted Erlang mixture (right) for the simulated generalized Pareto data.

6 Discussion

We extend the Lee and Lin (2010) EM algorithm for fitting mixtures of Erlangs with a common scale parameter to censored and truncated data. The EM algorithm able to deal with censored and truncated data remains a simple iterative algorithm. The initialization of the parameters can be done in a similar way as in Lee and Lin (2010) based on the denseness property (Tijms, 1994, p. 163) and provides close starting values making the algorithm converge fast. The shape adjustment procedure explores the parameter space in a clever way such that, when adjusting and reducing the shapes, the previous estimates for the scale and the weights provide a very close approximation to the maximum likelihood estimates corresponding to the new set of shapes,

which greatly reduces the run time. Extending Lee and Lin (2010), we suggest the use of a spread factor to achieve a wider spread for the shapes at the initial step. We recommend comparing the resulting fits starting from different initial values obtained by varying the spread factor and changing the initial number of Erlangs.

We implement the fitting procedure in `R` and show how mixtures of Erlangs can be used to adequately represent any univariate distribution in a wide variety of applications where data is allowed to be censored and truncated. We focus in the paper on the domain of actuarial science, where claim severity data is often censored and truncated due to policy modifications such as deductibles and policy limits. The use of mixtures of Erlangs offers on the one hand the flexibility of nonparametric density estimation techniques to describe the insurance losses and on the other hand the feasibility to analytically quantify the risk. The examples on several simulated and real datasets illustrate the effectiveness of our proposed algorithm and demonstrate the approximation strength of mixtures of Erlangs.

Future research may explore incorporating regressor variables in the mixture of Erlangs with common scale and introducing the flexibility of this approach in a regression context. We detected some limitations of mixtures of Erlangs in approximating heavy-tailed distributions and suggest combining our methodology with the extreme value methodology using a body-tail approach (Lee et al., 2012; Pigeon and Denuit, 2011). Adjusting the EM algorithm tailored to the class of multivariate mixtures of Erlangs, introduced by Lee and Lin (2012), to the case of censored and truncated data is another appealing extension.

Acknowledgements

The authors are grateful to the referees, to the editor and to Gerda Claeskens and Simon Lee for valuable comments and suggestions. Andrei Badescu and Sheldon Lin acknowledge the financial support received from the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was further supported in by NWO (Katrien Antonio, through a Veni 2009 grant), FWO (Katrien Antonio) and IWT (Roel Verbelen).

Appendix A Denseness

Theorem A.1. *(Tijms, 1994, p. 163) The class of mixtures of Erlang distributions with a common scale parameter is dense in the space of distributions on $\mathbb{R}^+$. More specifically, let $F(x)$ be the cumulative distribution function of a positive random variable. Define the following cumulative distribution function of a mixture of Erlang distributions with a common scale parameter $\theta > 0$,*

$$F(x;\theta) = \sum_{j=1}^{\infty} \alpha_j(\theta) F(x;j,\theta)\,,$$

where $F(x;j,\theta)$ denotes the cumulative distribution function of an Erlang distribution with shape j and scale θ,

$$F(x;j,\theta) = 1 - \sum_{n=0}^{j-1} e^{-x/\theta} \frac{(x/\theta)^n}{n!}\,,$$

and the mixing weights are given by

$$\alpha_j(\theta) = F(j\theta) - F((j-1)\theta) \qquad \text{for } j = 1, 2, \ldots .$$

Then

$$\lim_{\theta \to 0} F(x;\theta) = F(x)\,,$$

for each point x at which $F(\cdot)$ is continuous.

Appendix B Partial derivative of Q

We first introduce the lower incomplete gamma function

$$\gamma(s,x) = \int_0^x z^{s-1} e^{-z} dz\,,$$

by which we can write the cumulative distribution function of an Erlang distribution as

$$F(x;r,\theta) = \int_0^x \frac{z^{r-1}e^{-z/\theta}}{\theta^r (r-1)!} dz = \frac{1}{(r-1)!}\int_0^{x/\theta} u^{r-1}e^{-u} du = \frac{\gamma(r, x/\theta)}{(r-1)!}\,. \tag{22}$$

In order to maximize $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ with respect to θ, we set the first order partial derivative at $\theta^{(k)}$ equal to zero

$$\begin{aligned}
&\left.\frac{\partial Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})}{\partial \theta}\right|_{\theta=\theta^{(k)}} \\
&\quad = \sum_{i\in U}\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \left(\frac{x_i}{\theta^2} - \frac{r_j}{\theta} - \frac{\frac{\partial}{\partial\theta}\left[F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\right]}{F(t^u; r_j, \theta) - F(t^l; r_j, \theta)} \right) \\
&\qquad \sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} \left(\frac{E\left(X_i \left| Z_{ij} = 1, l_i, u_i, t^l, t^u; \theta^{(k-1)}\right.\right)}{\theta^2} - \frac{r_j}{\theta} \right. \\
&\qquad \left. - \frac{\frac{\partial}{\partial\theta}\left[F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\right]}{F(t^u; r_j, \theta) - F(t^l; r_j, \theta)} \right)\Bigg|_{\theta=\theta^{(k)}} \\
&\overset{(22)}{=} \frac{1}{\theta^2}\sum_{i\in U}\left(\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)}\right) x_i + \frac{1}{\theta^2}\sum_{i\in C}\left(\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} E\left(X_i \left| Z_{ij} = 1, l_i, u_i, t^l, t^u; \theta^{(k-1)}\right.\right)\right) \\
&\qquad - \frac{n}{\theta}\sum_{j=1}^{M}\left(\frac{\sum_{i\in U} {}^{u}z_{ij}^{(k)} + \sum_{i\in C} {}^{c}z_{ij}^{(k)}}{n}\right) r_j \\
&\qquad - \sum_{i\in U}\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \frac{\frac{\partial}{\partial\theta}\left(\gamma(r_j, t^u/\theta) - \gamma(r_j, t^l/\theta)\right)}{(r_j-1)!\left(F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\right)} \\
&\qquad - \sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} \frac{\frac{\partial}{\partial\theta}\left(\gamma(r_j, t^u/\theta) - \gamma(r_j, t^l/\theta)\right)}{(r_j-1)!\left(F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\right)}\Bigg|_{\theta=\theta^{(k)}}
\end{aligned}$$

$$
\begin{aligned}
\overset{(16)}{=} & \frac{1}{\theta^2}\sum_{i\in U} x_i + \frac{1}{\theta^2}\sum_{i\in C} E\left(X_i \middle| l_i, u_i, t^l, t^u; \theta^{(k-1)}\right) - \frac{n}{\theta}\sum_{j=1}^{M}\beta_j^{(k)} r_j \\
& - \sum_{i\in U}\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \frac{t^l/\theta^2\left(t^l/\theta\right)^{r_j-1} e^{-t^l/\theta} - t^u/\theta^2\left(t^u/\theta\right)^{r_j-1} e^{-t^u/\theta}}{(r_j-1)!\left(F(t^u;r_j,\theta) - F(t^l;r_j,\theta)\right)} \\
& - \sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} \frac{t^l/\theta^2\left(t^l/\theta\right)^{r_j-1} e^{-t^l/\theta} - t^u/\theta^2\left(t^u/\theta\right)^{r_j-1} e^{-t^u/\theta}}{(r_j-1)!\left(F(t^u;r_j,\theta) - F(t^l;r_j,\theta)\right)}\Bigg|_{\theta=\theta^{(k)}} \\
= & \frac{1}{\theta^2}\sum_{i\in U} x_i + \frac{1}{\theta^2}\sum_{i\in C} E\left(X_i \middle| l_i, u_i, t^l, t^u; \theta^{(k-1)}\right) - \frac{n}{\theta}\sum_{j=1}^{M}\beta_j^{(k)} r_j \\
& - \frac{n}{\theta^2}\sum_{j=1}^{M}\beta_j^{(k)} \frac{\left(t^l\right)^{r_j} e^{-t^l/\theta} - \left(t^u\right)^{r_j} e^{-t^u/\theta}}{\theta^{r_j-1}(r_j-1)!\left(F(t^u;r_j,\theta) - F(t^l;r_j,\theta)\right)}\Bigg|_{\theta=\theta^{(k)}} = 0\,,
\end{aligned}
$$

where we used expression (22) of the cumulative distribution of an Erlang.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Antonio, K. and Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 2014(7):649–669.

Badescu, A., Gong, L., Lin, X. S., and Tang, D. (2015). Modeling correlated frequencies with applications in operational risk management. *Journal of Operational Risk*, 10(1):1–43.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D., and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley.

Bolancé, C., Guillén, M., Gustafsson, J., and Nielsen, J. P. (2012). *Quantitative operational risk models*. CRC Press.

Cameron, A. and Trivedi, P. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Chernobai, A., Rachev, S., and Fabozzi, F. (2014). Composite goodness-of-fit tests for left-truncated loss samples. In Lee, C.-F. and Lee, J. C., editors, *Handbook of Financial Econometrics and Statistics*, pages 575–596. Springer New York.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dufour, R. and Maag, U. (1978). Distribution results for modified kolmogorov-smirnov statistics for truncated or censored. *Technometrics*, 20(1):29–32.

Frees, E. W. and Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, 103(484):1457–1469.

Guilbaud, O. (1988). Exact kolmogorov-type tests for left-truncated and/or right-censored data. *Journal of the American Statistical Association*, 83(401):213–221.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag, Heidelberg, second edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Klugman, S. and Rioux, J. (2006). Toward a unified approach to fitting loss models. *North American Actuarial Journal*, 10(1):63–83.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions*, volume 715. Wiley.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2013). *Loss models: Further topics*. John Wiley & Sons.

Lee, D., Li, W. K., and Wong, T. S. T. (2012). Modeling insurance claims via a mixture exponential model combined with peaks-over-threshold approach. *Insurance: Mathematics and Economics*, 51(3):538 – 550.

Lee, G. and Scott, C. (2012). EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816 – 2829.

Lee, S. C. and Lin, X. S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, 14(1):107–130.

Lee, S. C. and Lin, X. S. (2012). Modeling dependent risks with multivariate Erlang mixtures. *ASTIN Bulletin*, 42(1):153–180.

McCall, B. P. (1996). Unemployment insurance rules, joblessness, and part-time work. *Econometrica*, 64(3):647–82.

McLachlan, G. and Jones, P. (1988). Fitting mixture models to grouped and truncated data via the em algorithm. *Biometrics*, pages 571–578.

McLachlan, G. and Peel, D. (2001). *Finite mixture models*. Wiley.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*, volume 382. Wiley-Interscience.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. The John Hopkins University Press.

Pigeon, M. and Denuit, M. (2011). Composite lognormal-Pareto model with random threshold. *Scandinavian Actuarial Journal*, 2011(3):177–192.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Tijms, H. C. (1994). *Stochastic models: an algorithmic approach*. Wiley.

Willmot, G. E. and Lin, X. S. (2011). Risk modelling with the mixed Erlang distribution. *Applied Stochastic Models in Business and Industry*, 27(1):2–16.

Willmot, G. E. and Woo, J.-K. (2007). On the class of Erlang mixtures with risk theoretic applications. *North American Actuarial Journal*, 11(2):99–115.