KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Multivariate mixtures of Erlangs for density estimation under censoring and truncation: additional examples

Verbelen R, Antonio K, Claeskens G.

KBI_1511

Multivariate mixtures of Erlangs for density estimation under censoring and truncation: additional examples

Roel Verbelen*[1], Katrien Antonio[1,2], and Gerda Claeskens[1]

[1]LStat, Faculty of Economics and Business, KU Leuven, Belgium.
[2]Faculty of Economics and Business, University of Amsterdam, The Netherlands.

March 31, 2015

Abstract

In this addendum to Verbelen et al. (2015), we present several additional examples of the calibration procedure for fitting multivariate mixtures of Erlangs to censored and truncated data.

1 Mastitis study

An alternative mastitis dataset than the one being considered in Verbelen et al. (2015), contains information, for a total of 1196 cows, on the time until infection of the udder quarters by one specific bacterium, Corynebacterium bovis (CBO), rather than the time to infection by any bacteria. The smaller dataset used in the paper is not a subset of this alternative version. In the same way as we did in the paper, we label the dimensions as in Table 1 to indicate the udder quarters and estimate the underlying density using MME based on this four-dimensional sample of interval- and right-censored udder infection times.

Table 1: Labeling of the udder quarters.

Udder quarter	
RL	Rear Left
FL	Front Left
RR	Rear Right
FR	Front Right

A similar search for the tuning parameters as before, revealed that the same choice of setting $M = 20$ and $s = 10$ results in the best-fitting MME which this time contains five shape parameter combinations (Table 2).

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

Table 2: Parameter estimates of the best-fitting MME with four mixture components fitted to the mastitis data (infections by all bacteria).

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(2, 2, 2, 2)	0.2081	54.8592
(4, 3, 4, 3)	0.1712	
(4, 8, 3, 7)	0.0566	
(7, 5, 7, 6)	0.1673	
(24, 24, 11, 11)	0.3968	

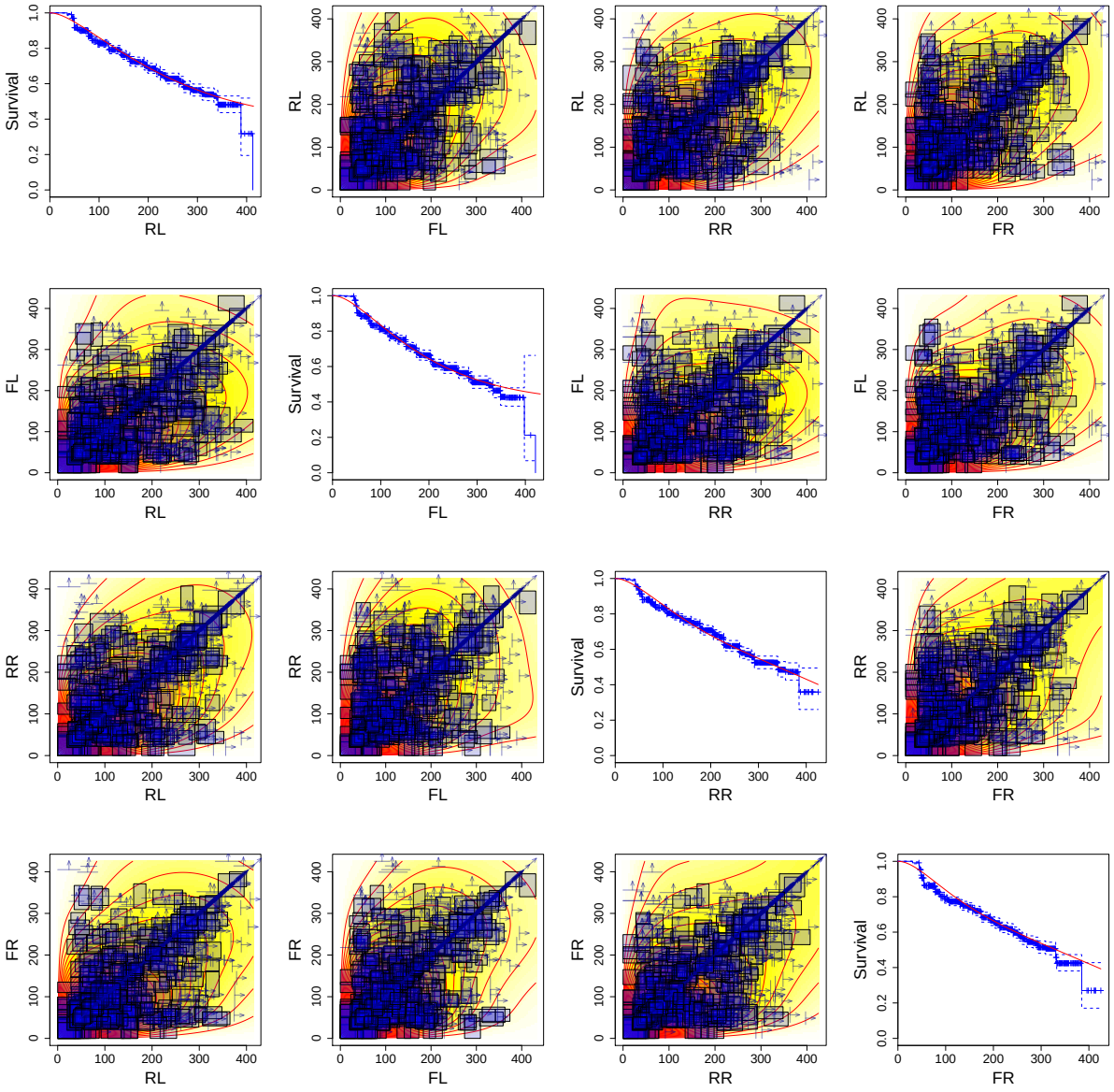

Figure 1: Scatterplot matrix comparing the fitted four-dimensional MME to the observed interval and right censored observations of the mastitis data (infections by CBO).

In Figure 1, we construct the scatterplot matrix to graphically assess the goodness-of-fit. The marginal survival functions from the MME smooth the nonparametric Turnbull estimate and the bivariate graphs show that the dependence is well represented. The estimates for Kendall's τ and Spearman's ρ, based on the fitted MME are reported in Table 3.

Table 3: Estimates for the bivariate measures of association Kendall's τ and Spearman's ρ based on the fitted MME for the mastitis data (infections by CBO).

		RL	FL	RR
FL	τ	0.5727		
	ρ	0.8030		
RR	τ	0.5530	0.5060	
	ρ	0.7729	0.7244	
FR	τ	0.5426	0.5494	0.4874
	ρ	0.7613	0.7712	0.6969

2 Danish fire insurance data

We consider the Danish insurance dataset containing information on 2167 fire losses over the period 1980 to 1990 of which the total loss exceeds 1 million Danish Krone. The data have been adjusted for inflation to reflect 1985 values and are expressed in millions of Danish Krone. The total loss amount is subdivided in damage to building, damage to content (e.g. furniture and personal property) and loss of profits. These data were collected at the Copenhagen Reinsurance Company and used for example in McNeil (1997); Embrechts et al. (1997); Drees and Müller (2008). This data set is available at `www.ma.hw.ac.uk/~mcneil/`.

We will estimate the density of two subsets of this dataset using MME. The data are three dimensional and consist of loss to building, loss to contents and loss to profits. First, we consider only the 517 claims for which the loss to each component has a non-zero value (separate, two-dimensional models would be needed for cases with loss to buildings and content but not to profits and so on). We fit a three dimensional MME to this subset and perform a grid search for the tuning parameters of the fitting procedure where we set M equal to 20, 100 and 517 (i.e. the sample size resulting in the maximal number of marginal quantiles) and let s vary between 100 and 1000 by 100 and between 1500 and 3000 by 500. The values of BIC of the resulting MME are plotted in Figure 2. The best fit is obtained for $M = 517$ and $s = 400$. The best-fitting MME contains 21 mixture components of which the parameter estimates are given in Table 4. A graphical goodness-of-fit scatterplot is constructed in Figure 3. Since the data are heavy tailed, we construct our plots on the log scales. On the diagonal, we construct histograms of the logarithm of the marginal losses along with the log transformed marginal fitted densities. On the off-diagonal, we construct bivariate scatterplots on the log scale with an overlay of the fitted bivariate density of the best-fitting MME using a contour plot and heat map. Overall, the fit appears to be good, capturing both the marginals as well as the dependence structure appropriately. On the log scale, it is however visually clear that MME are not able to extrapolate the heaviness in the tail. The bivariate densities on the off-diagonal are not as smooth in the upper right corners as they are in the body of the distribution. In the univariate case, the same remark was made (Verbelen et al., 2014, Example 5.4).

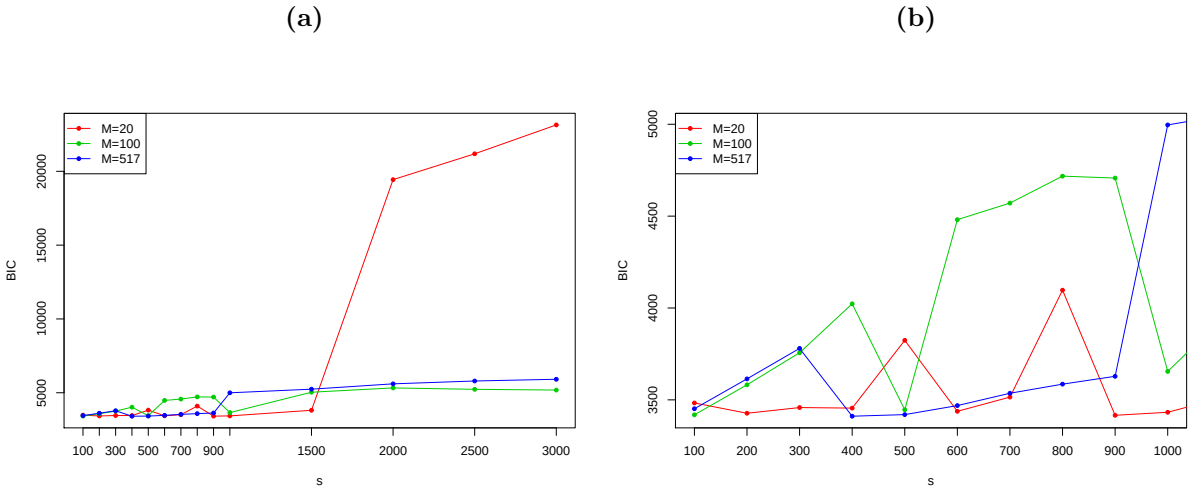

Figure 2: In (a), we show BIC values when fitting an MME to the Danish fire insurance data, starting from different values of the tuning parameters. In (b), we zoom in for s in between 100 and 1000. The minimum BIC value is obtained for M = 517 and s = 400.

Table 4: Parameter estimates of the best-fitting MME with 21 mixture components fitted to the Danish fire insurance data.

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(5, 2, 1)	0.5934	0.2477
(1, 4, 2)	0.0716	
(17, 27, 5)	0.0201	
(13, 4, 2)	0.0919	
(4, 9, 4)	0.0816	
(153, 49, 6)	0.0039	
(39, 8, 3)	0.0080	
(43, 17, 15)	0.0077	
(7, 15, 3)	0.0548	
(6, 11, 22)	0.0058	
(6, 45, 16)	0.0154	
(26, 41, 14)	0.0089	
(22, 69, 11)	0.0118	
(48, 533, 4)	0.0019	
(41, 158, 13)	0.0039	
(21, 8, 72)	0.0019	
(14, 32, 27)	0.0039	
(36, 60, 25)	0.0039	
(50, 30, 39)	0.0019	
(20, 86, 43)	0.0058	
(384, 429, 250)	0.0019	

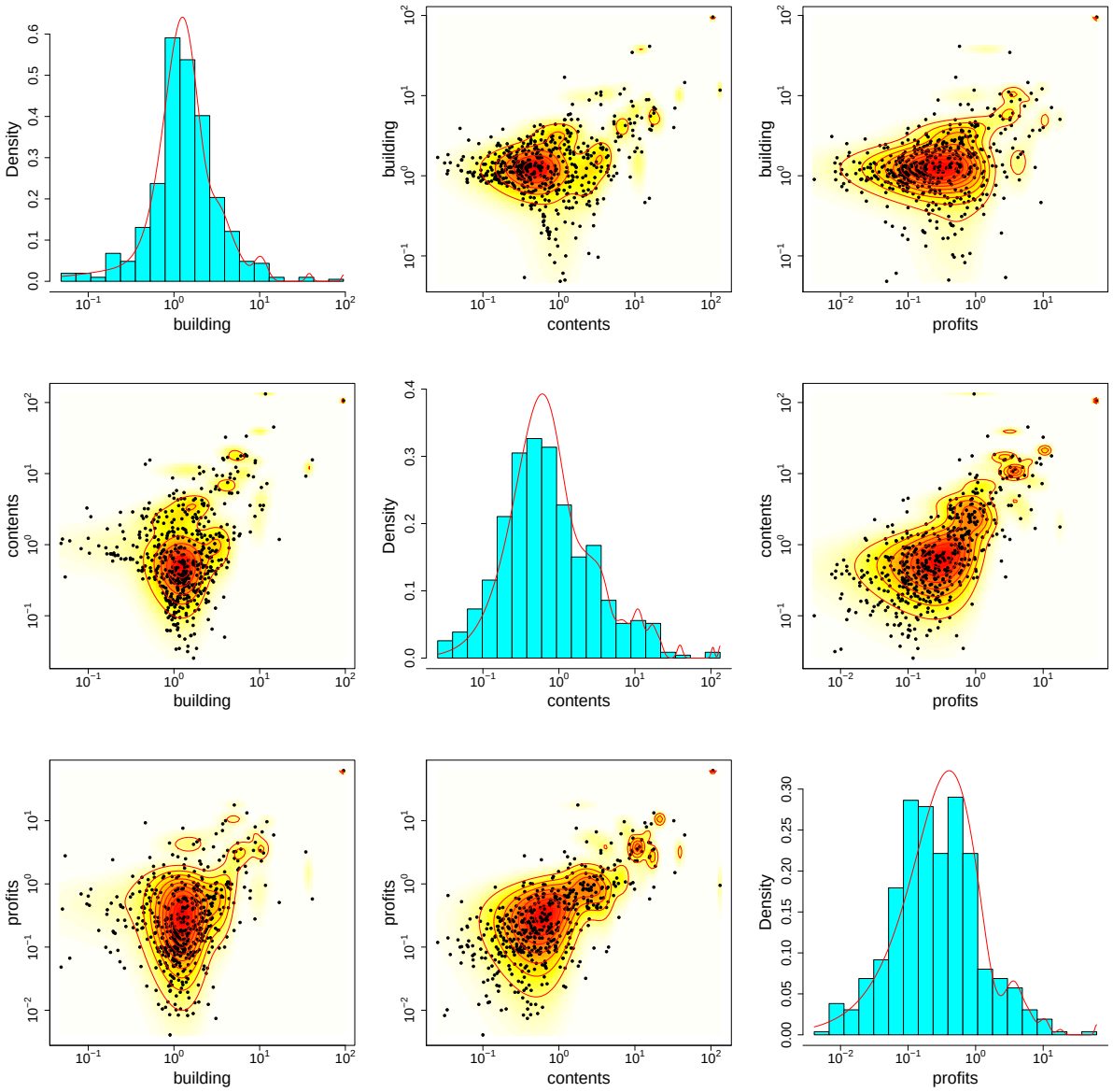

Figure 3: Scatterplot matrix comparing the fitted MME to the Danish fire insurance data on the log scale.

The total loss amount is the sum of the loss to building, loss to contents and loss to profits. By fitting a three dimensional MME to the components of the total loss, we can immediately also derive the distribution of the total loss amount itself, which is a univariate mixture of Erlangs. The graphical comparison of the fitted density to the empirical histogram of the total losses on the log scales reveals the same difficulty of smoothly fitting the tail.

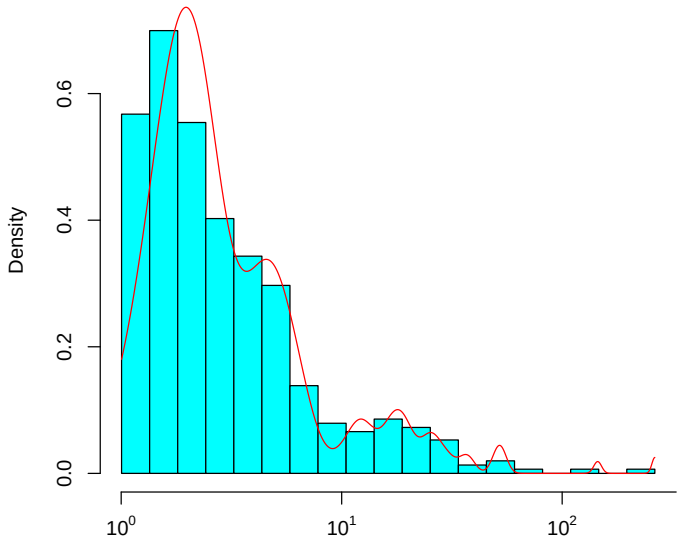

Figure 4: Graphical comparison of the fitted density of the sum of the components of the fitted three dimensional MME and the histogram of the observed total loss amounts of the Danish fire insurance data on the log scale.

When fitting an MME to the subset of strictly positive data, we neglected the fact that the data are in fact truncated. Indeed, a claim is only registered if the total loss exceeds 1 million Danish Krone, i.e. if $X_1 + X_2 + X_3 \geqslant 1$ where X_1 represents the loss to building, X_2 to contents and X_3 to profits. As a result, the fitted distribution of the sum of all components allows for a positive probability of 0.04 that we observe a claim for which the total loss is smaller than 1 (or smaller than 0 on the log scale), whereas this probability should be equal to zero.

This kind of censoring can however not be taken into account using the extended fitting procedure of Verbelen et al. (2015). The truncation region has to be of the rectangular form $[\boldsymbol{t}^l, \boldsymbol{t}^u]$. To illustrate the estimation under such kind of truncation, we take a subset of the data. In Drees and Müller (2008); Haug et al. (2011), they bypass the truncation issue by considering the two dimensional subset of losses to building and contents for which each component itself is larger than 1 million Danish Krone. We do the same, leading to a subset of 301 data points.

We fit a three dimensional MME to the data, taking the truncation into account, for different values of the tuning parameters. We set M equal to 10, 20 and 301 and let s vary between 10 and 100 by 10 and between 200 and 1000 by 100. The parameter estimates of the best fit according to BIC, obtained with tuning parameter $M = 10$ and $s = 500$, are shown in Table 5. A graphical goodness-of-fit plot for the three dimensional data is constructed in Figure 5.

Table 5: Parameter estimates of the best-fitting MME with 7 mixture components fitted to the truncated Danish fire insurance data.

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(1, 1)	0.9007	1.529272
(8, 2)	0.02734	
(63, 70)	0.0011	
(3, 25)	0.0080	
(25, 8)	0.0023	
(3, 7)	0.0595	
(8, 87)	0.0011	

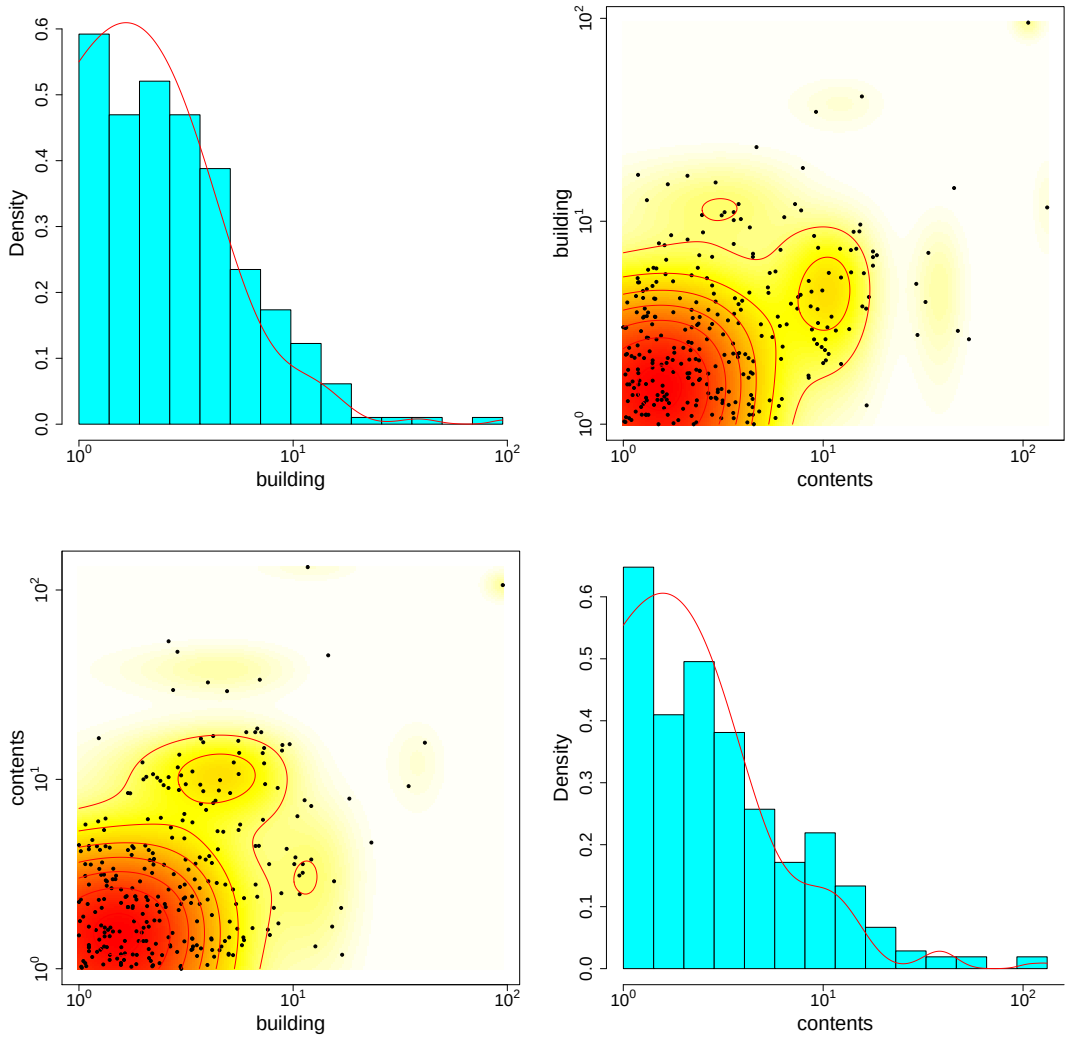

Figure 5: Scatterplot matrix comparing the fitted MME to the truncated two-dimensional Danish fire insurance data on the log scale.

3 Loss ALAE data

We consider an insurance dataset, collected by the US Insurance Services Office (ISO), comprising of 1500 non-life insurance claims of which both the indemnity payment or loss as well as the allocated loss adjustment expense (ALAE) are observed, both in USD. ALAE is the additional expense associated with the settlement of the claim, e.g. lawyers' fees , experts' opinions, and claims investigation expenses. For each claim, we also recorded the policy limit of the contract, i.e. the maximal claim amount. Due to the policy limits, 34 claims have a loss which is right censored as the amount of the claim exceeded the policy limit (a common feature of loss data). Even though only 34 of the 1500 observations are right censored, the censoring cannot be neglected and has to be taken into account when estimating the joint distribution. For instance, the mean of the censored losses is much higher than the mean of the uncensored losses (USD 217 941.2 versus USD 37 109.58). A peculiarity of the data is that there are a lot of ties, probably due to monetary rounding and precision issues. The loss amounts and ALAEs only consist of 542 and 1433 unique values. This dataset is also studied in e.g. Frees and Valdez (1998); Klugman and Parsa (1999); Beirlant et al. (2004); Denuit et al. (2006a,b).

3.1 Fitting a MME on the original scale

The objective is to describe the joint distribution of the losses and the expenses for which we use MME. The right censoring is taken into account when fitting the MME to the data. We perform a grid search to find good values for the tuning parameters. We set M equal to 20, 100 and 1500 and let s vary between 10 and 100 by 10 and between 200 and 1000 by 100. $M = 20$ and $s = 50$ turn out to be the best values for the tuning parameters in this grid. In Table 6, the parameter estimates of the fitted MME are given. Based on Figure 6, the marginals as well as the dependence structure seems to be captured appropriately, but again we notice the lack of smoothness of the bivariate fitted MME density in the upper right corner of the scatterplot, indicating the difficulty to extrapolate the heaviness of the tail.

Table 6: Parameter estimates of the best-fitting MME with 14 mixture components fitted to the loss ALAE data.

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(1, 1)	0.5585	6797.973
(3, 1)	0.1181	
(9, 1)	0.0461	
(4, 10)	0.0175	
(17, 17)	0.0070	
(33, 1)	0.0111	
(5, 2)	0.1237	
(15, 2)	0.0680	
(59, 5)	0.0131	
(9, 74)	0.0007	
(64, 32)	0.0053	
(31, 6)	0.0220	
(120, 1)	0.0042	
(320, 10)	0.0047	

From Figure 6, it is clear that there is a positive dependence between the amount of the loss and the ALAE: expensive claims tend to be associated with large settlement costs. Based on the fitted MME, which takes the right censoring into account, the bivariate measures of association Kendall's τ and Spearman's ρ get estimated as 0.2172 and 0.3203. For comparison, in Denuit et al. (2006b) an estimate for Kendall's τ equal to 0.3669 is obtained by assuming an Archimedian copula and using a nonparametric estimator of the joint distribution taking censoring into account. In Frees and Valdez (1998), the data are modeled using Pareto marginals and Gumbel-Hougaard copula, taking the right censoring of the losses into account, leading to corresponding estimate of 0.31 for Spearman's ρ, with 95% confidence interval of $(0.28, 0.34)$.

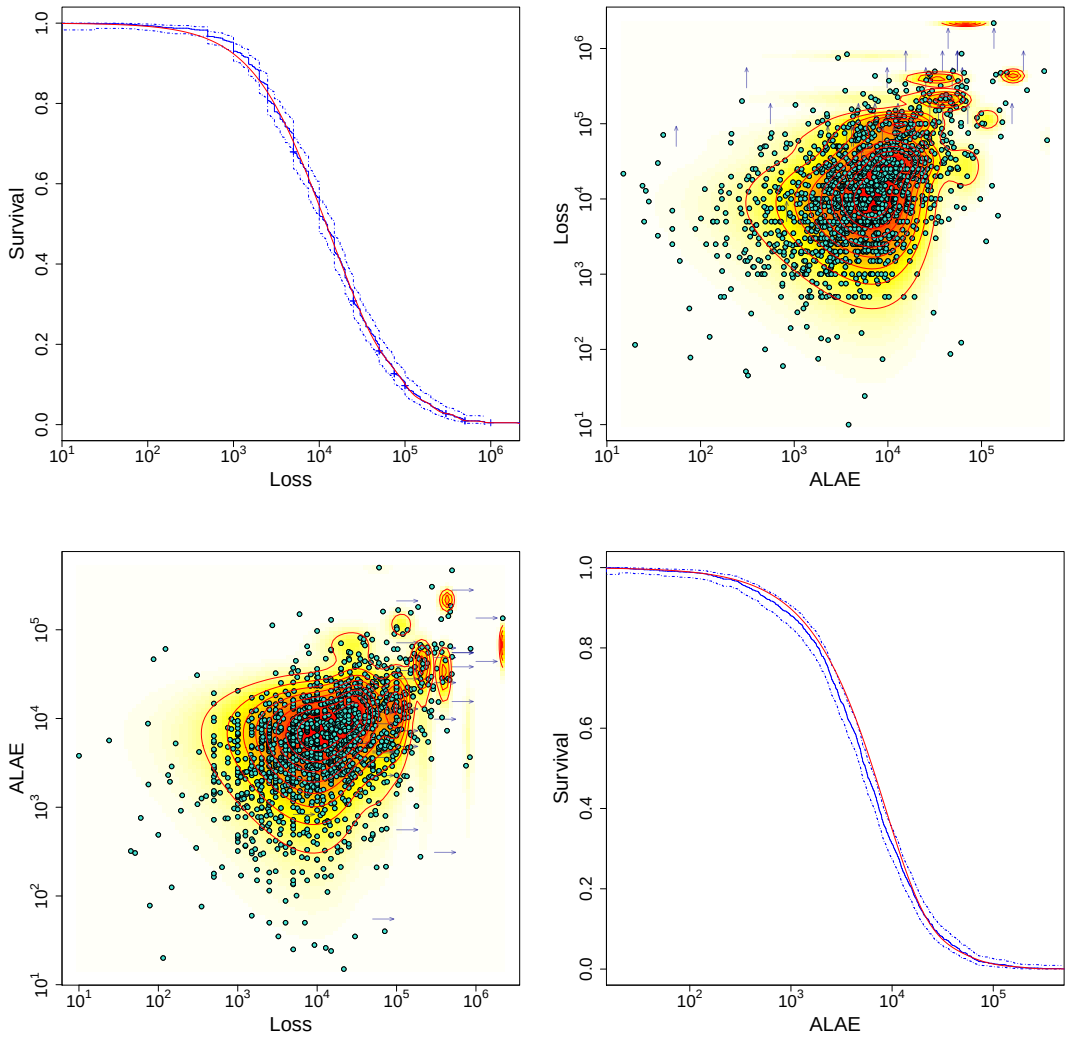

Figure 6: Scatterplot matrix comparing the fitted MME to the loss ALAE data on the log scale.

Estimating regression functions As in Frees and Valdez (1998); Klugman and Parsa (1999); Denuit et al. (2006a), we now consider the conditional distribution of ALAE given the value of the claim loss. Denoting the loss amount and the ALAE of a claim by (X, Y), this means we look at the conditional distribution $F_Y(y|X = x)$. It is easy to see that for MME, conditional

distributions again belong to the MME class and the values of the parameters are easily determined. In our case, the conditional distribution of ALAE given the loss is a univariate mixture of Erlangs. Hence, the conditional distribution is completely known and we can for instance estimate the conditional expected value and quantiles of the ALAE for a given value of the loss. In Figure 7, we display how the regression curves of the conditional expected value and of several quantiles of the ALAE fluctuate for different conditioning values of the loss. Based on one of these regression curves, an estimated capital can be preallocated for the ALAE in case the loss amount is known.

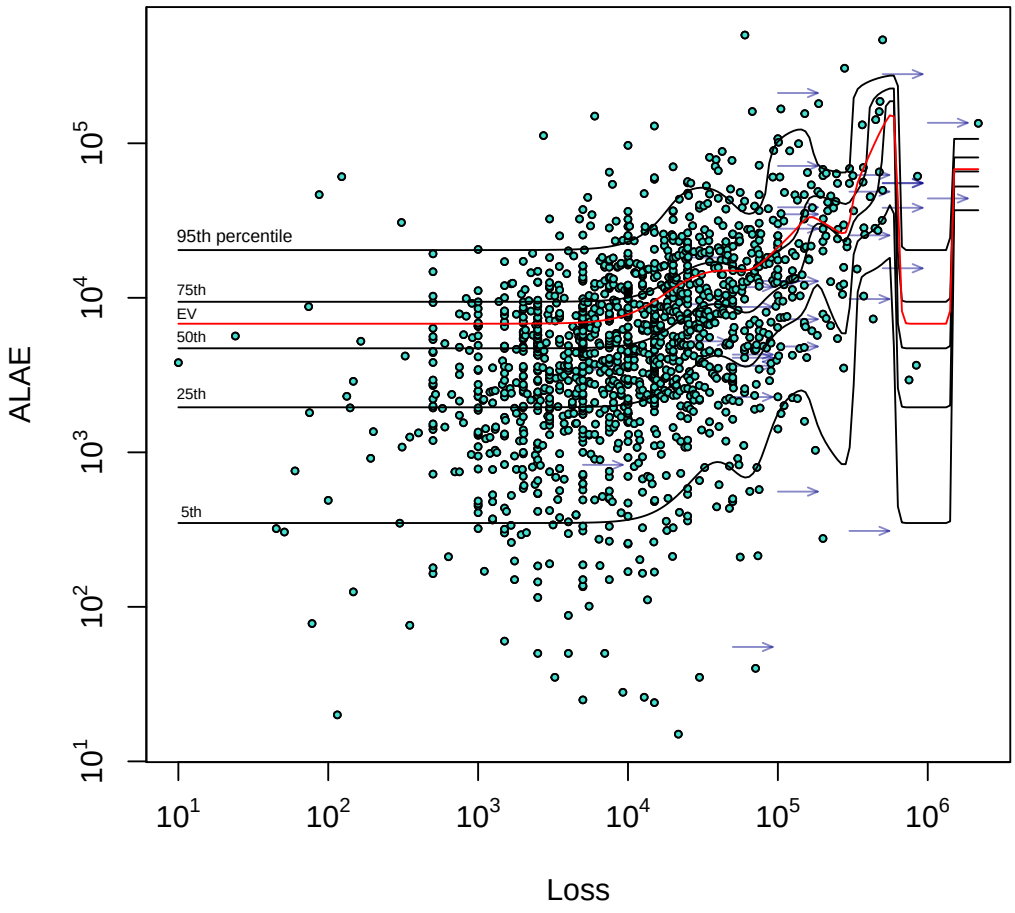

Figure 7: Scatterplot of the loss ALAE data on the log scale with regression functions for the expected value (EV) and the 5, 25, 50, 75 and 95 quantiles of ALAE based on the fitted MME.

Total claim cost The total claim cost for the insurer is the sum of the loss amount and the ALAE. The fitted MME on the bivariate data leads to a fitted univariate mixture of Erlangs distribution for the sum which takes the dependence between the losses and the expenses into account. In Figure 8, we compare the fitted survival function of the sum to the nonparametric

Kaplan-Meier estimate along with 95% confidence bounds. The '+' signs indicate the locations of the right censored total amounts. Based on the fitted univariate mixture of Erlangs distribution of the sum, we can for instance easily compute several risk measure using analytical formulas of univariate mixture of Erlangs. In Table 7, we show the values of the Value-at-Risk (VaR), Tail-Value-at-Risk (TVaR) and Expected Shortfall (ESF) for various levels of the confidence.

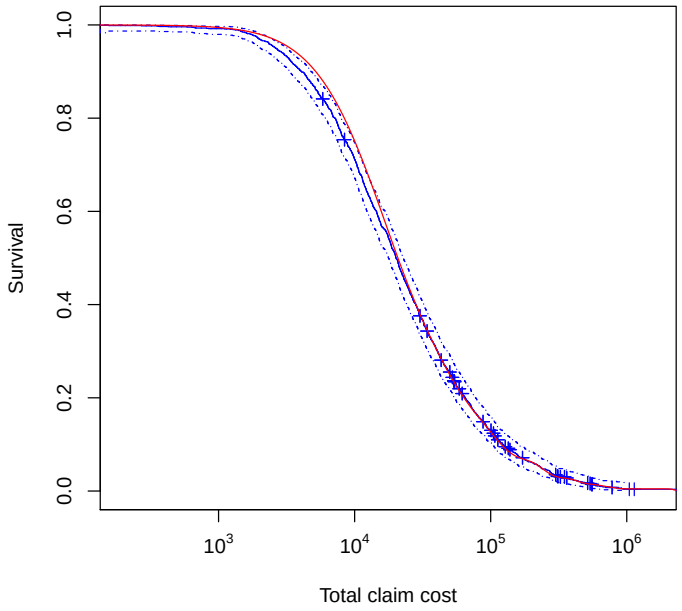

Figure 8: Graphical comparison of the fitted survival curve of the sum of the loss and ALAE, based on the fitted MME, and the Kaplan-Meier estimate, along with 95% confidence bounds, for the loss ALAE data on the log scale.

Table 7: Values of the Value-at-Risk (VaR), Tail-Value-at-Risk (TVaR) and Expected Shortfall (ESF) for various levels of the confidence of the sum for the loss ALAE data based on the fitted MME.

Confidence level	VaR	TVaR	ESF
0.80	63 609	232 557	33 790
0.85	86 038	285 443	29 911
0.90	119 572	377 329	25 776
0.95	234 864	589 995	17 757
0.99	698 886	1 479 901	7810

Calculating reinsurance premiums Reinsurers can use the fitted distribution of the sum, which takes the dependence between losses and ALAE into account, to price an excess-of-loss reinsurance layer. In this type of reinsurance, the reinsurer pays the part $f(Z)$ of the total claim

amount $Z = X + Y$ that exceeds a certain retention R, limited to a policy limit L:

$$f(Z) = \begin{cases} 0 & Z < R \\ Z - R & R \leqslant Z < L \\ L - R & X \geqslant L\,. \end{cases}$$

Such an excess-of-loss reinsurance layer is denoted by C xs R, where $C = L - R$ denotes the maximal cover amount. The net premium of this treaty has an explicit expression if Z has a univariate mixture of Erlangs distribution, as is the case here. In Table 8, the net reinsurance premium of this excess-of-loss reinsurance layer C xs R for various values of the retention R and the limit L are given.

Table 8: Values of the net reinsurance premium of an excess-of-loss reinsurance layer C xs R for various levels of the retention R and the limit L based on the fitted MME to the loss ALAE data.

	Ratio of the retention to the policy limit (R/L)				
Policy limit (L)	0.00	0.25	0.50	0.75	0.95
10 000	8963	6490	4141	1973	379
100 000	34 581	17 031	8842	3699	654
500 000	52 138	14 823	6618	2636	425
1 000 000	56 743	11 222	4605	1514	240

Typically however, the retention and policy limit of an excess-of-loss reinsurance treaty only apply to the claim losses. The associated settlement costs are shared on a pro-rata basis (cfr. Frees and Valdez, 1998). Again, it is crucial to account for the dependence between the loss and the ALAE to avoid an underestimation of the payment for the reinsurer. More specifically, again denoting the retention by R and the limit by L, the reinsurer's payment for a given realization (X, Y) of a loss and an associated ALAE is described by the function

$$g(X,Y) = \begin{cases} 0 & X < R \\ X - R + \dfrac{X-R}{X} Y & R \leqslant X < L \\ L - R + \dfrac{L-R}{L} Y & X \geqslant L\,. \end{cases}$$

Even for MME, the net premium $E(g(X,Y))$ of such a reinsurance treaty does not have an explicit expression and requires simulation (as is done using copulas in Frees and Valdez (1998)) or numerical integration. Having an appropriate model available for the joint distribution of the loss ALAE couple (X, Y) is essential. For MME, generating a random sample is simple: first sample a mixture component using the weights as probabilities and then generate an observation from the corresponding joint distribution of independent Erlangs. Based on N sampled couples (X_i, Y_i), the net premium $E(g(X,Y))$ gets estimated as

$$\widehat{E}(g(X,Y)) = \frac{1}{N} \sum_{i=1}^{N} g(X_i, Y_i)$$

with standard error

$$SE(\widehat{E}(g(X,Y))) = \sqrt{\frac{\frac{1}{N-1}\sum_{i=1}^{N}[g(X_i,Y_i) - \widehat{E}(g(X,Y))]^2}{N}} = \sqrt{\frac{\frac{1}{N-1}\sum_{i=1}^{N} g(X_i,Y_i)^2 - \widehat{E}(g(X,Y))^2}{N}}\,.$$

Simulated values of the net reinsurance premium of this kind of excess-of-loss reinsurance layer with pro rata sharing of expenses based on $N = 1\,000\,000$ simulations are shown in Table 9 for various levels of the retention R and the limit L. The corresponding standard errors are denoted between brackets.

Table 9: Simulated values of the net reinsurance premium of an excess-of-loss reinsurance layer C xs R with pro rata sharing of expenses for various levels of the retention R and the limit L based on the fitted MME to the loss ALAE data. The simulated values are based on $N = 1\,000\,000$ simulations and the standard errors are denoted between brackets.

	Ratio of the retention to the policy limit (R/L)				
Policy limit (L)	0.00	0.25	0.50	0.75	0.95
10 000	20 057(27)	13 207(21)	8236(14)	3921(7)	759(1)
100 000	39 450(48)	17 591(37)	9438(24)	4075(12)	726(2)
500 000	53 448(95)	13 618(63)	6090(37)	2145(16)	292(3)
1 000 000	57 166(120)	9897(72)	3895(42)	1566(19)	255(4)

3.2 Fitting a MME on the log scale

Alternative, in order to get a smoother fit in the tail, one can fit the data on the log scale, as is also done in e.g. Di Bernardino and Rullière (2013). Hence, we take the (natural) logarithm of the data before fitting a MME.

The best values for the tuning parameters of the fitting procedure for MME are $M = 20$ and $s = 110$ when comparing the resulting fits of a grid search over M equal to 10 or 20 and s varying between 10 and 200 by 10 and between 300 and 1000 by 100. The parameter estimates of the corresponding fitted MME are given in Tabel 10. Figure 9 evaluates the fit on the log scale. The fitted marginal survival functions almost perfectly coincide with the Kaplan-Meier estimates and bivariate fitted density closely agrees with the scatterplot and attains a smoother fit in the tail.

The downside to this approach is the fact that we lose most of the tractability of working with MME. The fitted distribution on the log scale is a MME, but the same does not hold on the original scale. Hence, in the following we mostly have to rely on Monte Carlo simulation to evaluate the fitted distribution on the original scale.

Table 10: Parameter estimates of the best-fitting MME with 8 mixture components fitted to the log transformed loss ALAE data.

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(95, 37)	0.0087	0.0976
(81, 61)	0.0468	
(101, 65)	0.0423	
(48, 50)	0.0049	
(46, 85)	0.0090	
(81, 83)	0.2709	
(101, 91)	0.4969	
(121, 105)	0.1205	

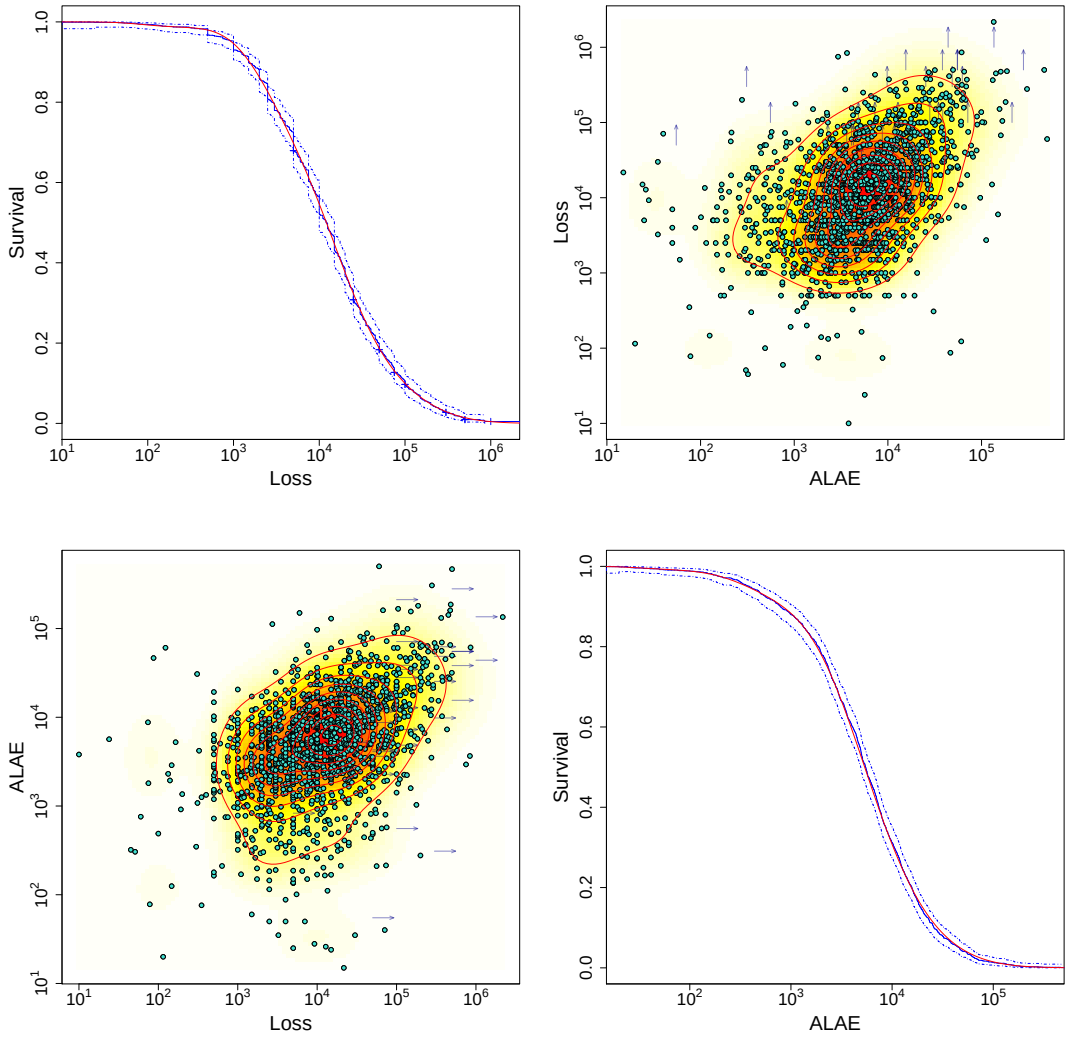

Figure 9: Scatterplot matrix comparing the fitted MME to the log transformed loss ALAE data.

Estimating regression functions Since quantiles on the log scale can be immediately transformed to quantiles on the original scale by taking the exponential, quantile regression can be done analytically. The same does not hold for the expected value. Therefore, for each value of the conditioning loss variable, we generate a sample $N = 1\,000\,000$ observations from the conditional marginal univariate mixture of Erlangs and take the mean of the exponentiated observations as an estimate for the expected value on the original scale. The resulting regression curves are shown in Figure 10 which are not as erratic in the tail as was the case in Figure 7.

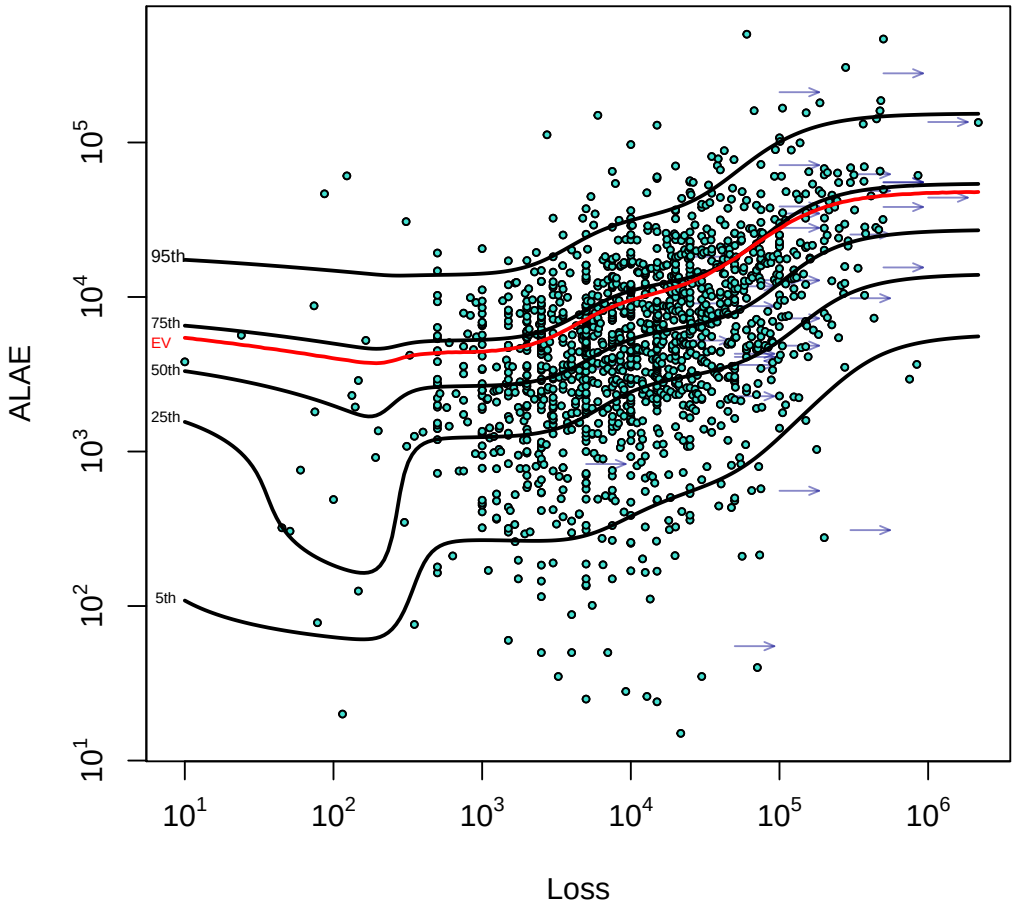

Figure 10: Scatterplot of loss ALAE data on the log scale with regression functions for the expected value on the original scale (EV) and the 5, 25, 50, 75 and 95 quantiles of ALAE based on the fitted MME to the log transformed data. The values for the expected value are simulated based on $N = 1\,000\,000$ simulations of the conditional marginal univariate mixture of Erlangs.

Total claim cost In order to evaluate the fit of the sum of the loss and ALAE component of the claim, we consider the values of the Value-at-Risk (VaR), the Tail-Value-at-Risk (TVaR) and Expected Shortfall (ESF) of the total claim cost for various levels of the confidence, as we did before. These values are computed using the empirical versions of these risk measures based on a Monte Carlo sample of $N = 1\,000\,000$ simulated observations of the fitted bivariate MME on the log scale. The simulated values are shown in Table 11 and quite closely follow the values in Table 7 obtained from the fitted MME on the original scale. Standard errors are included between brackets for the TVaR and ESF estimates.

Table 11: Simulated values of the Value-at-Risk (VaR), the Tail-Value-at-Risk (TVaR) and Expected Shortfall (ESF) for various levels of the confidence of the sum for the loss ALAE data based on the fitted MME on the log transformed data. The simulated values are based on $N = 1\,000\,000$ simulations and the standard errors are denoted between brackets for the TVaR and ESF.

Confidence level	VaR	TVaR	ESF
0.80	63 970	231 679(923)	33 542(185)
0.85	86 727	284 151(1209)	29 614(181)
0.90	130 744	373 225(1755)	24 248(175)
0.95	241 301	570 687(3269)	16 469(163)
0.99	690 853	1 316 988(13395)	6261(134)

Calculating reinsurance premiums Based on the same Monte Carlo sample of $N = 1\,000\,000$ simulated observations of the fitted bivariate MME on the log scale, we now compute the simulated estimates for the excess-of-loss reinsurance layer on the total cost of the claim in Table 12 and on the loss of the claim with pro rata sharing of expenses in Table 13. Standard errors are denoted between brackets. The simulated net reinsurance premiums predominantly have the same order of magnitude as the previously obtained values in Tables 8 and 9.

Table 12: Simulated values of the net reinsurance premium of an excess-of-loss reinsurance layer C xs R for various levels of the retention R and the limit L based on the fitted MME to the log transformed loss ALAE data. The simulated values are based on $N = 1\,000\,000$ simulations and the standard errors are denoted between brackets.

	Ratio of the retention to the policy limit (R/L)				
Policy limit (L)	0.00	0.25	0.50	0.75	0.95
10 000	8651(2)	6187(2)	3904(2)	1855(1)	358(1)
100 000	34 269(33)	17 199(27)	8911(18)	3741(9)	670(2)
500 000	53 244(92)	16 055(63)	7263(38)	2725(18)	453(3)
1 000 000	57 964(120)	11 984(75)	4721(43)	1652(19)	265(4)

Table 13: Simulated values of the net reinsurance premium of an excess-of-loss reinsurance layer C xs R with pro rata sharing of expenses for various levels of the retention R and the limit L based on the fitted MME to the log transformed loss ALAE data. The simulated values are based on $N = 1\,000\,000$ simulations and the standard errors are denoted between brackets.

	Ratio of the retention to the policy limit (R/L)				
Policy limit (L)	0.00	0.25	0.50	0.75	0.95
10 000	20 296(33)	13 661(25)	8546(17)	4071(8)	788(2)
100 000	39 463(51)	17 535(38)	9211(25)	3917(12)	707(2)
500 000	54 246(98)	14 064(64)	6359(38)	2419(18)	407(3)
1 000 000	58 263(123)	10 483(73)	4231(42)	1508(19)	243(4)

4 Claims development data

We consider an insurance claims dataset from a European insurance company. We want to investigate and model the dependence between the development delay, i.e. the time in between the reporting and settlement of a claim (expressed in days), and the total cost of a claim. We focus on the information on bodily injury claims between January 1997 and August 2009 for which the initial reserve was smaller than 10 000. The analysis date is September 1, 2009. Claims which are not settled at the time of analysis lead to right censored observations for the development delay as well as for the total cost, which exists of the sum of the loss payments paid thus far, discounted to January 1, 1997 using the consumer price inflation index. The dataset consists of 4492 claims of which 261 were right censored at the time of the analysis. A more detailed version of this dataset is analyzed in Antonio and Plat (2014).

Setting the tuning parameters equal to $M = 20$ and $s = 400$ is considered to be a good choice after a comparing the resulting fits using BIC obtained with $M = 20$ and letting s vary between 10 and 100 by 10 and between 200 and 1000 by 100 and with $M = 100$ and letting s vary between 10 and 100 by 10. Parameter estimates are shown in Table 14 and a graphical evaluation of the fit is made in Figure 11. Both the development delay and the total payment are skewed to the right and heavy tailed, which is why we evaluate the fit on the log scale. Note that the the right censored observations cluster to the right in the lower left plot in Figure 11: the mean development delay and total payment of the settled claims equals 423 and 3173, respectively, whereas the mean of the right censoring points for the development delay and the total payment of the open claims equals 1432 and 13 890, respectively. Our model represents the distribution of both the development delay and the total payment well and captures the positive dependence in the data. Kendall's τ and Spearman's ρ of the fitted MME are given by 0.4382 and 0.6094, respectively. In Figure 12, you can see how the conditional quantiles of the total payment change along with the conditioning values of the development delay.

Table 14: Parameter estimates of the best-fitting MME with 26 mixture components fitted to the claims development data.

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(1, 2)	0.4383	186.632
(11, 2)	0.0116	
(1, 16)	0.0125	
(2, 7)	0.2135	
(12, 13)	0.0073	
(2, 13)	0.0316	
(3, 16)	0.0715	
(10, 29)	0.0132	
(3, 27)	0.0532	
(4, 41)	0.0342	
(4, 55)	0.0150	
(5, 79)	0.0156	
(2, 97)	0.0012	
(10, 63)	0.0057	
(6, 106)	0.0133	
(16, 103)	0.0024	
(6, 138)	0.0074	
(16, 143)	0.0027	
(7, 182)	0.0056	
(7, 219)	0.0043	
(9, 289)	0.0027	
(9, 377)	0.0030	
(13, 646)	0.0048	
(6, 991)	0.0006	
(15, 1807)	0.0009	
(400, 30246)	0.0279	

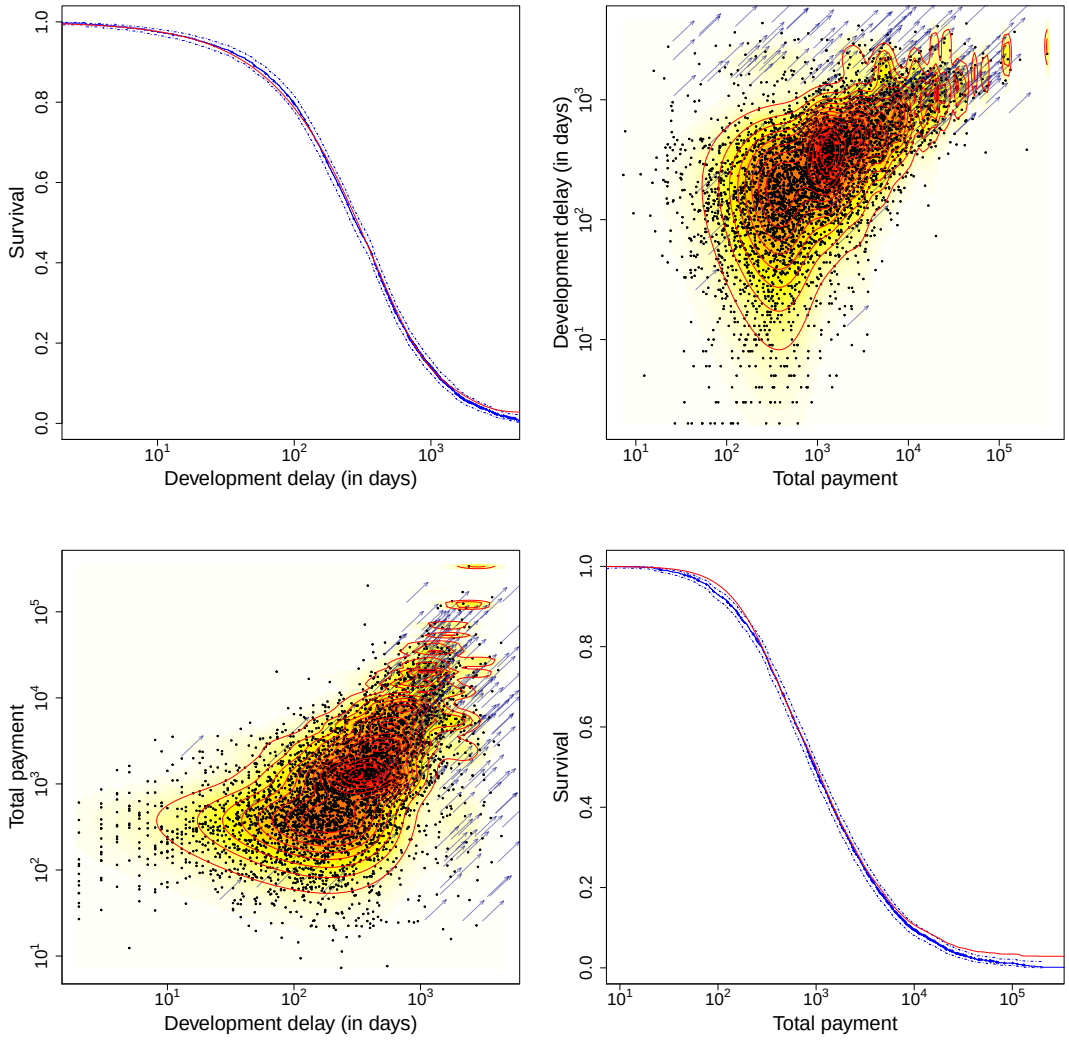

Figure 11: Scatterplot matrix comparing the fitted MME to the claims development data on the log scale.

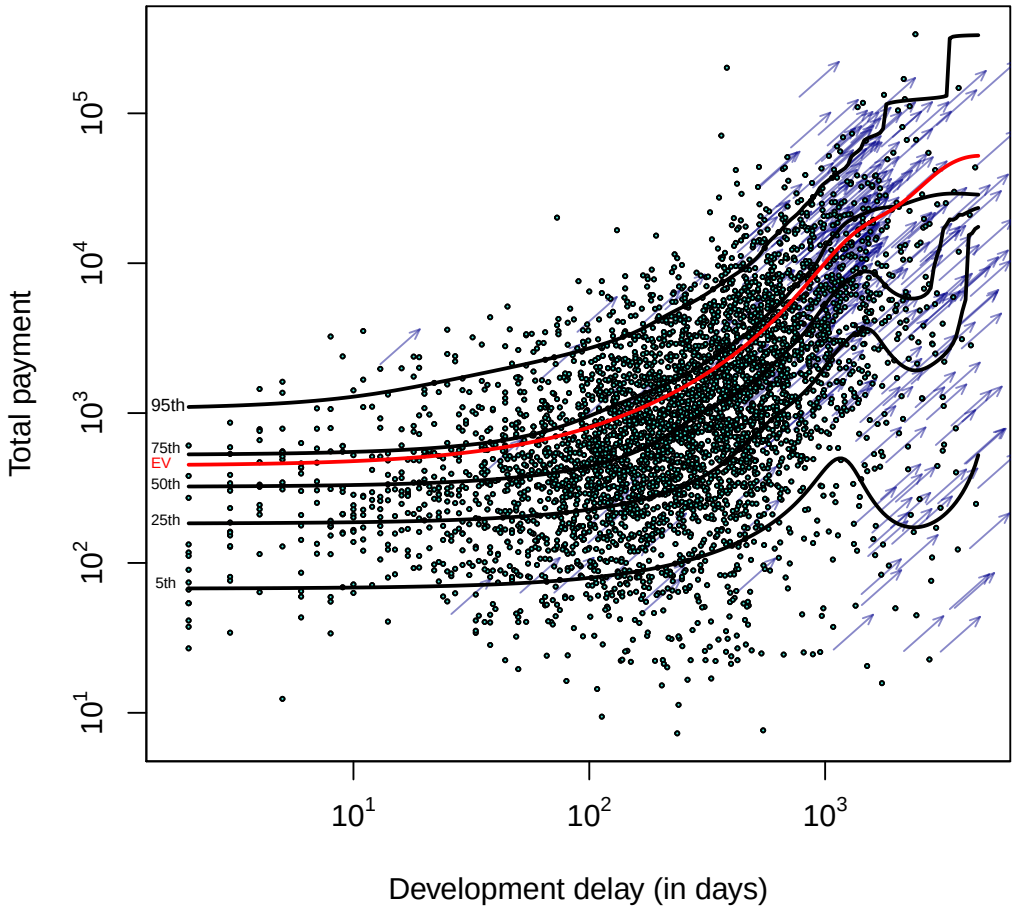

Figure 12: Scatterplot of the claims development data on the log scale with regression functions for the expected value (EV) and the 5, 25, 50, 75 and 95 quantiles of ALAE based on the fitted MME.

References

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

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.

Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. (2006a). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Wiley.

Denuit, M., Purcaru, O., and Van Keilegom, I. (2006b). Bivariate archimedean copula models for censored data in non-life insurance. *Journal of Actuarial Practice*, 13:5–32.

Di Bernardino, E. and Rullière, D. (2013). On certain transformation of Archimedean copulas: Application to the non-parametric estimation of their generators. *Dependence Modeling*, 1:Pages 1–36, ISSN (Online) 2300–2298.

Drees, H. and Müller, P. (2008). Fitting and validation of a bivariate model for large claims. *Insurance: Mathematics and Economics*, 42(2):638 – 650.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events: For Insurance and Finance*. Applications of mathematics. Springer.

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American actuarial journal*, 2(1):1–25.

Haug, S., Klüppelberg, C., and Peng, L. (2011). Statistical models and methods for dependence in insurance data. *Journal of the Korean Statistical Society*, 40(2):125 – 139.

Klugman, S. A. and Parsa, R. (1999). Fitting bivariate loss distributions with copulas. *Insurance: Mathematics and Economics*, 24(1-2):139–148.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, 27:117–137.

Verbelen, R., Antonio, K., and Claeskens, G. (2015). Multivariate mixtures of Erlangs for density estimation under censoring and truncation. *Submitted for publication*.

Verbelen, R., Gong, L., Antonio, K., Badescu, A., and Lin, X. S. (2014). Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *Submitted for publication*.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be