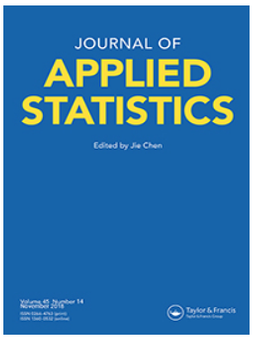

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

# Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions

Antonio Punzo, Angelo Mazza & Antonello Maruotti

**To cite this article:** Antonio Punzo, Angelo Mazza & Antonello Maruotti (2018) Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions, Journal of Applied Statistics, 45:14, 2563-2584, DOI: 10.1080/02664763.2018.1428288

**To link to this article:** https://doi.org/10.1080/02664763.2018.1428288

JOURNAL OF APPLIED STATISTICS, 2018
VOL. 45, NO. 14, 2563–2584
https://doi.org/10.1080/02664763.2018.1428288



# Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions

Antonio Punzo [a], Angelo Mazza [a] and Antonello Maruotti [b,c]

[a]Dipartimento di Economia e Impresa, Università di Catania, Catania, Italy; [b]Dipartimento di Scienze Economiche, Politiche e delle Lingue Moderne, Libera Università Maria Ss. Assunta, Roma, Italy; [c]Centre for Innovation and Leadership in Health Sciences, University of Southampton, Southampton, UK

**ABSTRACT**

Insurance and economic data are frequently characterized by positivity, skewness, leptokurtosis, and multi-modality; although many parametric models have been used in the literature, often these peculiarities call for more flexible approaches. Here, we propose a finite mixture of contaminated gamma distributions that provides a better characterization of data. It is placed in between parametric and non-parametric density estimation and strikes a balance between these alternatives, as a large class of densities can be implemented. We adopt a maximum likelihood approach to estimate the model parameters, providing the likelihood and the expected-maximization algorithm implemented to estimate all unknown parameters. We apply our approach to an artificial dataset and to two well-known datasets as the workers compensation data and the healthcare expenditure data taken from the medical expenditure panel survey. The Value-at-Risk is evaluated and comparisons with other benchmark models are provided.

**ARTICLE HISTORY**
Received 24 September 2016
Accepted 10 January 2018

**KEYWORDS**
Contaminated distributions; gamma distribution; finite mixtures; insurance claims; robust estimation

## 1. Introduction

Insurance and economic data are often positive, right-skewed, leptokurtic, and multi-modal [40]. Though many parametric models have been used in the actuarial and economic literature [16,43], the peculiarities of these data call for more flexible models.

Skewed distributions may be considered to accommodate right-skewness, and their need has been first emphasized by Lane [44] for insurance risk. In this class of distributions, Bodnar and Gupta [13] and Vernic [61] identify the skew-normal as a promising model in the context of financial asset returns and risk measurement, respectively. However, using the skew-normal distribution is in principle appropriate when the support is $\mathbb{R}$, while it is not adequate if the support is $\mathbb{R}^+$ as it causes *boundary bias*, that is, allocation of probability mass outside the theoretical support. A simple remedy is to use

**CONTACT** A. Maruotti a.maruotti@lumsa.it, a.maruotti@soton.ac.uk

distributions defined on $\mathbb{R}^+$. Following this idea, Azzalini *et al.* [5] consider the log-skew-normal distribution to model claims and household income data. Other classical examples are the gamma [57], the log-normal [9] and the Weibull [33] distributions, only to cite a few.

Whereas each of the distributions mentioned above assumes unimodality, insurance and economic data often show a behavior hardly compatible with the choice of fitting a single parametric distribution [15]. In such cases, a more flexible modeling framework is needed. Here, we focus on a finite mixture approach that shares the efficiency of parametric modeling and the flexibility of non-parametric density estimation techniques. The flexibility of finite mixtures in accommodating various shapes of the insurance and economic data is now widely recognized [11,18,39,48]. Among mixture models, mixtures of gamma distributions were successfully considered in [26,60,63] for insurance data. Bagnato and Punzo [6] focused on the subclass of mixtures of unimodal gamma densities, adopting a mode-based parameterization for the component distributions that leads to a sufficiently flexible approach.

Furthermore, as emphasized by Cowell and Victoria-Feser [21] in the case of economic data about incomes, and by Bulla [14], Choy *et al.* [19], and Punzo *et al.* [49] in the case of insurance and financial data, outliers – also referred to as 'bad' points herein, following Aitkin and Wilson [3] – 'contaminate' our dataset affecting, from an inferential point of view, the estimation of the parameters for the chosen model. Thus, outliers detection and the development of robust methods of parameter estimation insensitive to their presence are important tasks. As suggested by Davies and Gather [24], to properly define an observation as an outlier, it is crucial to define a *reference* distribution that models the bulk of the data; accordingly, the region where the density of this reference distribution is lower identifies the so-called region of outliers. Here, we attempt to account for all the possible features of insurance and economic data by introducing mixtures of contaminated gamma distributions. As special cases of our model, we obtain the unimodal gamma density, the contaminated gamma density, and the mixture of unimodal gamma densities. With respect to the latter approach, mixtures of contaminated gamma distributions also allow for an automatic detection of local outliers. Indeed, the contaminated gamma model is a two-component mixture in which one of the components, with a large prior probability, represents the reference distribution, and the other, with a small prior probability, the same mode, and an inflated variability, represents the outliers. It represents a common and simple theoretical model for the occurrence of outliers and the two additional parameters, with respect to the parameters of the reference distribution, have a direct interpretation in terms of proportion of 'good' data and degree of contamination (a sort of measure of how different outliers are from the bulk of the good incomes). Advantageously, such a model also allows for automatic detection of outliers via a simple and natural procedure based on maximum *a posteriori* probabilities [47,51–54].

The paper is organized as follows. Section 2 introduces the proposed model, sufficient conditions for its identifiability are given in Section 3, and an EM algorithm to obtain maximum likelihood parameters' estimates is illustrated in Section 4. Further aspects, related to the robustness of the proposed model, are discussed in Section 5. Artificial and real data are considered in Section 6 to appreciate the advantages of the proposed model. Section 7 summarizes the key aspects of the proposal along with future possible extensions.

## 2. Mixture of contaminated gamma distributions

The distribution of a random variable $X$, taking values on $\mathbb{R}^+$, according to a finite mixture model, can be written as

$$p(x;\boldsymbol{\psi}) = \sum_{j=1}^{k} \pi_j f\left(x;\boldsymbol{\vartheta}_j\right), \quad x > 0, \tag{1}$$

where $\pi_j$ is the mixing proportion for the $j$th component, with $\pi_j > 0$ and $\sum_{j=1}^{k} \pi_j = 1$, $f(x;\boldsymbol{\vartheta}_j)$ is the density of the $j$th component with parameters $\boldsymbol{\vartheta}_j$, and $\boldsymbol{\psi} = (\boldsymbol{\pi}', \boldsymbol{\vartheta}')'$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ and $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1', \ldots, \boldsymbol{\vartheta}_k')'$, contains all of the parameters of the mixture.

In this paper, for the $j$th mixture component, $j = 1, \ldots, k$, we adopt the contaminated gamma distribution

$$f\left(x;\boldsymbol{\vartheta}_j\right) = \alpha_j g\left(x;\lambda_j,\nu_j\right) + \left(1-\alpha_j\right) g\left(x;\lambda_j,\eta_j\nu_j\right), \quad x > 0, \tag{2}$$

where $\boldsymbol{\vartheta}_j = (\lambda_j, \nu_j, \alpha_j, \eta_j)'$ and

- $$g\left(x;\lambda_j,\nu_j\right) = \frac{x^{\lambda_j/\nu_j}\, \mathrm{e}^{-x/\nu_j}}{\nu_j^{\lambda_j/\nu_j+1}\,\Gamma\left(\frac{\lambda_j}{\nu_j}+1\right)}, \quad x > 0,$$

  is the mode-parametrized unimodal gamma density (see, e.g. [17]) chosen as reference distribution for $X$, with $\lambda_j > 0$ denoting the mode and $\nu_j > 0$ governing the concentration of $g$ around the mode. The adjective 'unimodal' is useful to highlight the subclass of gamma densities on which attention is focused on;
- $\alpha_j \in (0.5, 1)$ can be seen as the proportion of good data;
- $\eta_j > 1$ denotes the degree of contamination and, because of the assumption $\eta_j > 1$, it can be interpreted as the increase in variability due to the bad data with respect to the reference distribution $g(x;\lambda_j,\nu_j)$; hence, it is an inflation parameter.

The marginal effect of varying each parameter in $\boldsymbol{\vartheta}_j$, the other its components kept fixed, is illustrated by a set of contaminated gamma densities displayed in Figure 1. Of course, because both the reference distribution $g(x;\lambda_j,\nu_j)$ and the inflated distribution $g(x;\lambda_j,\eta_j\nu_j)$ have their maximum in $\lambda_j$, this also guarantees that $f$ will produce a unimodal density with mode $\lambda_j$. As a limiting case, when $\alpha_j \to 1^-$ and $\eta_j \to 1^+$, the reference distribution $g(x;\lambda_j,\nu_j)$ is obtained. Note that $\alpha_j$ is constrained to be greater than 0.5 because, in robust statistics, it is usually assumed that at least half of the points are good; however, $\alpha_j \in (0, 1)$ is acceptable in general, as often happens in the literature.

Substituting Equation (2) in Equation (1), the density of our mixture of contaminated gamma distributions is given by

$$p(x;\boldsymbol{\psi}) = \sum_{j=1}^{k} \pi_j \left[\alpha_j g\left(x;\lambda_j,\nu_j\right) + \left(1-\alpha_j\right) g\left(x;\lambda_j,\eta_j\nu_j\right)\right], \quad x > 0. \tag{3}$$

It easy to realize that there are $5k-1$ unknown parameters to be estimated. Note that, as a limiting case, when $\alpha_j \to 1^-$ and $\eta_j \to 1^+$ for $j = 1, \ldots, k$, the mixture of unimodal

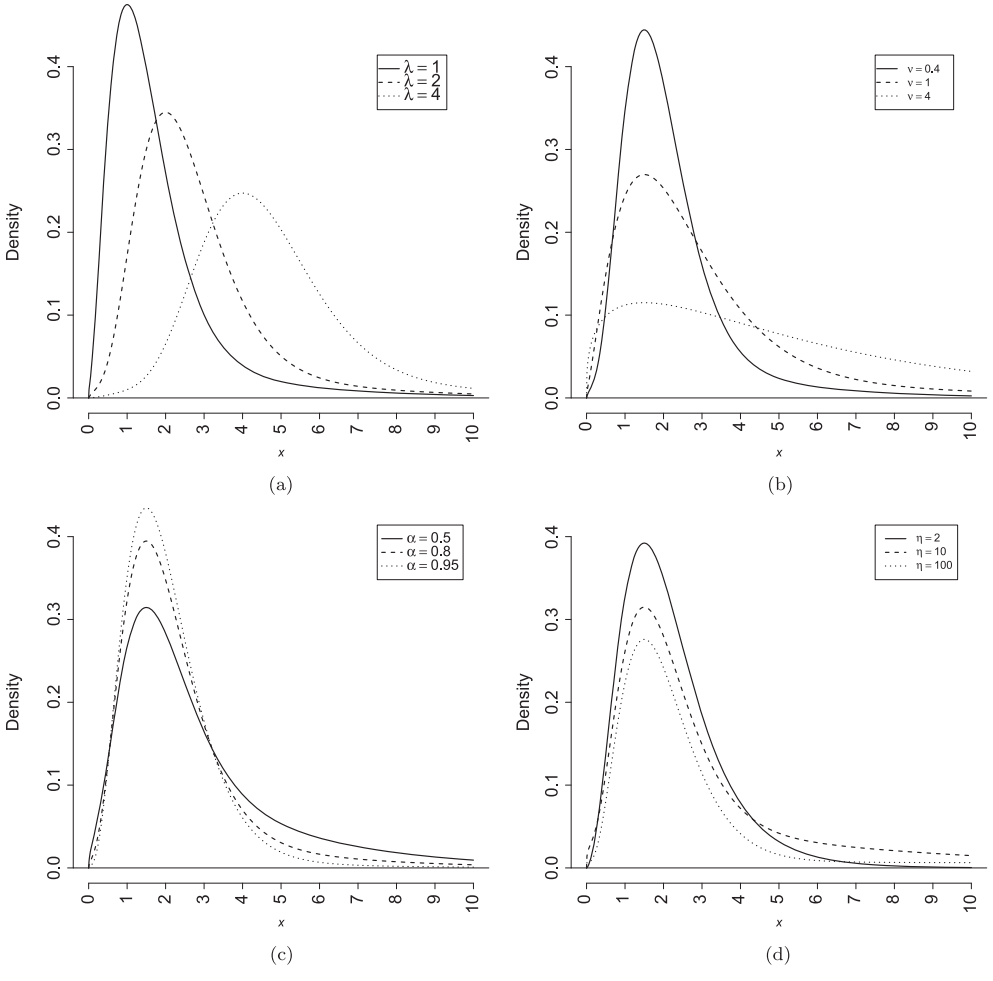


**Figure 1.** Contaminated gamma densities: (a) $\nu = 0.5$, $\alpha = 0.8$, and $\eta = 5$, (b) $\lambda = 1.5$, $\alpha = 0.8$, and $\eta = 5$, (c) $\lambda = 1.5$, $\nu = 0.5$, and $\eta = 5$, and (d) $\lambda = 1.5$, $\nu = 0.5$, and $\alpha = 0.6$.

gamma distributions defined in [6] is obtained. Of course, as also underlined by Izenman [41, p. 103], there is no guarantee that $p$ will produce a multimodal density with the same number of modes as there are densities in the mixture; similarly, there is no guarantee that those individual modes $\lambda_j$ will remain at the same locations in Equation (3). Indeed, the shape of the mixture distribution depends upon both the spacings of the modes and the relative shapes of the component distributions. Nevertheless, we maintain that the values of $\lambda_j$ should accurately approximate the location of the mixture-modes of $p(x; \boldsymbol{\psi})$ for well-separated components.

## 3. Identifiability

Before outlining parameter estimation for model (3), it is important to establish its identifiability. Identifiability is a necessary requirement, *inter alia*, for the usual asymptotic theory to hold for maximum likelihood (ML) estimation of the model parameters (cf. Section 4).

Before investigating identifiability, it is convenient to rewrite the model density as

$$p(x;\boldsymbol{\psi}) = \sum_{j=1}^{k}\sum_{h=1}^{2}\pi_j\alpha_{jh}g\left(x;\lambda_j,\eta_{jh}\nu_j\right), \tag{4}$$

where, with respect to Equation (3), $\alpha_{j1} = \alpha_j$, $\alpha_{j2} = 1 - \alpha_{j1}$, $\eta_{j1} = 1$, and $\eta_{j2} = \eta_j$.

A sufficient condition for the identifiability of finite mixtures of gamma distributions has been given by Teicher [59]. As stated by Di Zio *et al.* [27], in the absence of any constraint, a mixture of mixtures, such as model (4), is not identifiable in general; this is essentially due to the possibility of interchanging component labels between the two levels of the model.

In Proposition 3.1, it will be shown that our model is identifiable provided that, given two of the $k$ gamma distributions $g(x;\lambda_j,\nu_j)$ representing the good observations, they have distinct modes.

**Proposition 3.1:** *Let*

$$p(x;\boldsymbol{\psi}) = \sum_{j=1}^{k}\sum_{h=1}^{2}\pi_j\alpha_{jh}g\left(x;\lambda_j,\eta_{jh}\nu_j\right)$$

*and*

$$p\left(x;\tilde{\boldsymbol{\psi}}\right) = \sum_{s=1}^{\tilde{G}}\sum_{t=1}^{2}\tilde{\pi}_s\tilde{\alpha}_{st}g(x;\tilde{\lambda}_s,\tilde{\eta}_{st}\tilde{\nu}_s)$$

*be two different parameterizations of model* (4). *If* $j \neq j_1$ *implies*

$$\lambda_j \neq \lambda_{j_1}, \tag{5}$$

*then the equality* $p(x;\boldsymbol{\psi}) = p(x;\tilde{\boldsymbol{\psi}})$ *implies that* $k = \tilde{k}$ *and also implies that there exists a relabelling such that*

$$\pi_j = \tilde{\pi}_j, \quad \alpha_{jh} = \tilde{\alpha}_{jh}, \quad \lambda_j = \tilde{\lambda}_j, \quad \nu_j = \tilde{\nu}_j, \quad \text{and} \quad \eta_{jh} = \tilde{\eta}_{jh}.$$

***Proof:*** The identifiability of finite mixtures of gamma distributions guarantees that $2k = 2\tilde{k}$, i.e. $k = \tilde{k}$, and, for each pair $(j,h)$, there exists a pair $(s,t)$ such that

$$\pi_j\alpha_{jh} = \tilde{\pi}_s\tilde{\alpha}_{st}, \quad \lambda_j = \tilde{\lambda}_s, \quad \text{and} \quad \eta_{jh}\nu_j = \tilde{\eta}_{st}\tilde{\nu}_s; \tag{6}$$

cf. [27]. Note that Condition (5), the fact that $\eta_{j2} > \eta_{j1}$ ($\eta_{s2} > \eta_{s1}$), and the positivity of all the weights $\pi_j$ and $\alpha_{jh}$ ($\pi_s$ and $\alpha_{st}$) avoids nonidentifiability due to potential overfitting (a potential problem for identifiability first noted by [22]). In particular, the positivity constraint on the weights avoids nonidentifiability due to empty components while the remaining two constraints avoid nonidentifiability due to identical components.

Based on Condition (5), only two of the $2k$ gamma distributions – those with corresponding $g$ for the first parameterization ($s$ for the second) – can have the same mode.

Hence, for each pair $(j,s)$, with $j,s \in \{1,\ldots,k\}$, satisfying Equation (6), the problem reduces to comparing the pair

$$\left\{\left\{\pi_j\alpha_{j1}, \eta_{j1}\nu_j\right\}, \left\{\pi_j\alpha_{j2}, \eta_{j2}\nu_j\right\}\right\} \tag{7}$$

with the pair

$$\left\{\left\{\tilde{\pi}_s\tilde{\alpha}_{s1}, \tilde{\eta}_{s1}\tilde{\nu}_s\right\}, \left\{\tilde{\pi}_s\tilde{\alpha}_{s2}, \tilde{\eta}_{s2}\tilde{\nu}_s\right\}\right\}. \tag{8}$$

Thanks to the constraint that the inflation parameters $\eta_{j2}$ and $\eta_{s2}$ must be greater than one, it is easy to show that $\eta_{j2} = \tilde{\eta}_{s2}$ and $\nu_j = \tilde{\nu}_s$. In particular, if we compare the first variability term in Equation (7) with the first variability term in (8), and the second variability term in Equation (7) with the second variability term in Equation (8), we obtain

$$\begin{cases} \eta_{j1}\nu_j = \tilde{\eta}_{s1}\tilde{\nu}_s \\ \eta_{j2}\nu_j = \tilde{\eta}_{s2}\tilde{\nu}_s \end{cases} \Rightarrow \begin{cases} \eta_{j2} = \tilde{\eta}_{s2} \\ \nu_j = \tilde{\nu}_s \end{cases}, \tag{9}$$

which is exactly what we need for identifiability. In Equation (9), we have used the fact that, by definition, $\eta_{j1} = \tilde{\eta}_{s1} = 1$. On the contrary, if we consider the remaining possibility to compare the first variability term in Equation (7) with the second variability term in Equation (8), and the second variability term in Equation (7) with the first variability term in Equation (8), we obtain the impossible equation $\eta_{j2}\tilde{\eta}_{s2} = 1$; this equation is impossible because $\eta_{j2}$ and $\eta_{s2}$ are both greater than one.

With regard to the mixture weights, we know from Equation (9) that the first element of Equation (7) is related to the first element of Equation (8) and the second element of Equation (7) is related to the second element of Equation (8); accordingly, we have only to compare the corresponding weights. In particular, we obtain

$$\begin{aligned} &\begin{cases} \pi_j\alpha_{j1} = \tilde{\pi}_s\tilde{\alpha}_{s1} \\ \pi_j\alpha_{j2} = \tilde{\pi}_s\tilde{\alpha}_{s2} \end{cases} \Rightarrow \begin{cases} \pi_j\alpha_{j1} = \tilde{\pi}_s\tilde{\alpha}_{s1} \\ \pi_j\left(1-\alpha_{j1}\right) = \tilde{\pi}_s\left(1-\tilde{\alpha}_{s1}\right) \end{cases} \\ &\Rightarrow \begin{cases} \pi_j = \tilde{\pi}_s \\ \alpha_{j1} = \tilde{\alpha}_{s1} \end{cases}. \end{aligned} \tag{10}$$

Finally, based on Equations (6), (9), and (10), after a suitable relabelling, we obtain

$$\pi_j = \tilde{\pi}_j, \quad \alpha_{jh} = \tilde{\alpha}_{jh}, \quad \lambda_j = \tilde{\lambda}_j, \quad \nu_j = \tilde{\nu}_j, \quad \text{and} \quad \eta_{jh} = \tilde{\eta}_{jh},$$

with $j \in \{1,\ldots,k\}$ and $h \in \{1,2\}$, and this completes the proof. ■

## 4. Maximum likelihood estimation: the EM algorithm

To fit model (3), we use the expectation–maximization (EM) algorithm [25], which is a natural approach for ML estimation when data are incomplete. In our case, there are two sources of incompleteness. The first source, the classical one in the use of mixture models, arises from the fact that for each observation we do not know its component membership; this source is governed by an indicator vector $\boldsymbol{z}_i = (z_{i1},\ldots,z_{ik})$, where $z_{ij} = 1$ if $x_i$ comes from component $j$ and $z_{ij} = 0$ otherwise. The other source, which is specific for

this model, arises from the fact that for each observation we do not know if it is good or bad with reference to component $j$. To denote this source of incompleteness, we use $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ik})$, where $u_{ij} = 1$ if $x_i$ is good in component $j$ and $u_{ij} = 0$ if it is bad. The complete-data likelihood can be written as

$$L_c(\boldsymbol{\psi}) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left\{\pi_j\left[\alpha_j g\left(x_i;\lambda_j,\nu_j\right)\right]^{u_{ij}}\left[\left(1-\alpha_j\right)g\left(x_i;\lambda_j,\eta_j\nu_j\right)\right]^{1-u_{ij}}\right\}^{z_{ij}}. \tag{11}$$

Therefore, the complete-data log-likelihood, which is the core of the algorithm, becomes

$$l_c(\boldsymbol{\psi}) = l_{1c}(\boldsymbol{\pi}) + l_{2c}(\boldsymbol{\alpha}) + l_{3c}(\boldsymbol{\lambda},\boldsymbol{\nu},\boldsymbol{\eta}), \tag{12}$$

where $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_k)'$, $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_k)'$, $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_k)'$, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)'$, $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_k)'$,

$$l_{1c}(\boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\ln\pi_j, \tag{13}$$

$$l_{2c}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\left[u_{ij}\ln\alpha_j + \left(1-u_{ij}\right)\ln\left(1-\alpha_j\right)\right], \tag{14}$$

$$l_{3c}(\boldsymbol{\lambda},\boldsymbol{\nu},\boldsymbol{\eta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\left\{u_{ij}\ln\left[g\left(x_i;\lambda_j,\nu_j\right)\right] + \left(1-u_{ij}\right)\ln\left[g\left(x_i;\lambda_j,\eta_j\nu_j\right)\right]\right\}. \tag{15}$$

The EM algorithm iterates between two steps, an E-step and one M-step, until convergence. They are described below.

### 4.1. E-step

The E-step, on the $(r+1)$th iteration of the EM algorithm, requires the calculation of $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)})$, the current conditional expectation of $l_c(\boldsymbol{\psi})$. To do this, we need to calculate $E_{\boldsymbol{\psi}^{(r)}}(Z_{ij} \mid x_i)$ and $E_{\boldsymbol{\psi}^{(r)}}(U_{ij} \mid x_i, \boldsymbol{z}_i)$, $i = 1,\ldots,n$ and $j = 1,\ldots,k$. They are respectively given by

$$E_{\boldsymbol{\psi}^{(r)}}\left(Z_{ij} \mid x_i\right) = \frac{\pi_j^{(r)} f\left(x_i;\boldsymbol{\vartheta}_j^{(r)}\right)}{p\left(x_i;\boldsymbol{\psi}^{(r)}\right)} =: z_{ij}^{(r)}$$

and

$$E_{\boldsymbol{\psi}^{(r)}}\left(U_{ij} \mid x_i, Z_{ij}=1\right) = \frac{\alpha_j^{(r)} g\left(x_i;\lambda_j^{(r)},\nu_j^{(r)}\right)}{f\left(x_i;\boldsymbol{\vartheta}_j^{(r)}\right)} =: u_{ij}^{(r)}. \tag{16}$$

Then, by substituting $z_{ij}$ with $z_{ij}^{(r)}$ and $u_{ij}$ with $u_{ij}^{(r)}$ in Equation (12), and based on Equation (13)–(15), we obtain

$$Q\left(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)}\right) = Q_1\left(\boldsymbol{\pi} \mid \boldsymbol{\psi}^{(r)}\right) + Q_2\left(\boldsymbol{\alpha} \mid \boldsymbol{\psi}^{(r)}\right) + Q_3\left(\boldsymbol{\lambda},\boldsymbol{\nu},\boldsymbol{\eta} \mid \boldsymbol{\psi}^{(r)}\right). \tag{17}$$

### 4.2. M-step

The M-step on the $(r+1)$th iteration of the EM algorithm requires the calculation of $\boldsymbol{\psi}^{(r+1)}$ as the value of $\boldsymbol{\psi}$ that maximizes $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)})$. As the three terms on the right-hand side of Equation (17) have zero cross-derivatives, they can be maximized separately. Maximizing $Q_1(\boldsymbol{\pi} \mid \boldsymbol{\psi}^{(r)})$ with respect to $\boldsymbol{\pi}$, subject to the constraints on those parameters, yields

$$\pi_j^{(r+1)} = \frac{n_j^{(r)}}{n}, \quad j = 1, \ldots, k,$$

where $n_j^{(r)} = \sum_{i=1}^{n} z_{ij}^{(r)}$. Maximizing $Q_2(\boldsymbol{\alpha} \mid \boldsymbol{\psi}^{(r)})$ with respect to $\boldsymbol{\alpha}$, is equivalent to independently maximizing each of the $k$ expressions

$$Q_{2j}\left(\alpha_j \mid \boldsymbol{\psi}^{(r)}\right) = \sum_{i=1}^{n} z_{ij}^{(r)} \left[ u_{ij}^{(r)} \ln \alpha_j + \left(1 - u_{ij}^{(r)}\right) \ln\left(1 - \alpha_j\right) \right]$$

with respect to $\alpha_j$ subject to the constraint on this parameter, $j = 1, \ldots, k$. If we apply the constraint $\alpha_j \in (0, 1)$, then the solution is in closed-form and it is given by

$$\alpha_j^{(r+1)} = \frac{1}{n_j^{(r)}} \sum_{i=1}^{n} z_{ij}^{(r)} u_{ij}^{(r)}, \quad j = 1, \ldots, k, \tag{18}$$

If, based on the comments in Section 2, we apply the constraint $\alpha_j \in (0.5, 1)$, then the `optimize()` function, of the **stats** package for R [55], is used for a numerical search of the maximum $\alpha_j^{(r+1)}$ of $Q_{2j}(\alpha_j \mid \boldsymbol{\psi}^{(r)})$, over the interval $(0.5, 1)$. In the analyses herein, we use this approach to update $\alpha_j$.

Maximizing $Q_3(\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\eta} \mid \boldsymbol{\psi}^{(r)})$ with respect to $(\boldsymbol{\lambda}', \boldsymbol{\nu}', \boldsymbol{\eta}')'$ (subject to the constraints on these parameters) is equivalent to independently maximizing each of the $k$ expressions

$$\begin{aligned} Q_{3j}\left(\lambda_j, \nu_j, \eta_j \mid \boldsymbol{\psi}^{(r)}\right) = \sum_{i=1}^{n} z_{ij}^{(r)} \Big\{ & u_{ij}^{(r)} \ln\left[g\left(x_i; \lambda_j, \nu_j\right)\right] \\ & + \left(1 - u_{ij}^{(r)}\right) \ln\left[g\left(x_i; \lambda_j, \eta_j \nu_j\right)\right] \Big\} \end{aligned} \tag{19}$$

with respect to $(\lambda_j, \nu_j, \eta_j)' \in I\!R^+ \times I\!R^+ \times (1, \infty)$, $j = 1, \ldots, k$. Details on the partial derivatives of $Q_{3j}(\lambda_j, \nu_j, \eta_j \mid \boldsymbol{\psi}^{(r)})$ with respect to $\lambda_j$, $\nu_j$, and $\eta_j$, are given in the Appendix. The `optim()` function, of the **stats** package for R, can be used for a numerical search of the maximum. However, `optim()` does not allow for constrained maximization. To take into account the constraints on $\lambda_j$, $\nu_j$, and $\eta_j$, convenient monotonic transformations can be applied so to map these parameters to the real line. Back-transformations can be then adopted to obtain the updates $\lambda_j^{(r+1)}$, $\nu_j^{(r+1)}$, and $\eta_j^{(r+1)}$.

## 5. Robustness and treatment of outliers

### 5.1. Some notes about breakdown-robustness

As emphasized by Coretto and Hennig [20], although robustness results for some clustering methods can be found in the literature, robustness theory in cluster analysis remains

a tricky issue. In our model-based clustering setup, the mixture components are usually interpreted as clusters and the clusters are characterized by the parameters of their mixture components.

For a fixed number $k$ of components, Hennig [37] derived a robustness theory for mixture estimators based on the finite sample addition breakdown point by Donoho and Huber [28]. This breakdown point is defined, in general, as the smallest proportion of points that has to be added to a dataset in order to make the estimation arbitrarily bad, which is usually defined by at least one estimated component parameter converging to infinity (or zero for mixture proportions) under a sequence of a fixed number of added points. Hennig [37] showed that the addition breakdown point not only for mixtures of normal distributions, but also for mixtures of $t$ distributions, is the smallest possible; both these methods can be driven to a breakdown by adding a single data point. Note, however, that a point has to be a very extreme outlier (also said gross outlier) for the $t$-mixture to cause trouble, while it's much easier to drive conventional normal mixtures to a breakdown, in the sense that also mild outliers may cause problems for this model [56].

Under a mixture of $k$ unimodal gamma distributions, which is the model we assume for the good data (the so-called reference model; see [1,24,35,36]), the addition breakdown can defined as: $\lambda_j \to \infty$, $\nu_j \to \infty$, or $\pi_j \to 0$, for at least one of $j = 1, \ldots, k$ (see also [38]). Under our contaminated model (3), the tails-specific parameters $\alpha_j$ and $\eta_j$ are not regarded as interesting on their own, but as a helpful device to robustify the reference model and, as such, they are not included in the breakdown point definition. By analogy in terms of model's structure between the normal mixture and the unimodal gamma mixture, and between the $t$-mixture and the contaminated gamma mixture, due to the componentwise heavy tails of these models, we expect the same robustness results cited above: neither the unimodal gamma mixture nor the contaminated gamma mixture is theoretically breakdown robust, though in practice very extreme outliers are needed to spoil our method.

### 5.2. Automatic detection of bad points

For the proposed model, the classification of an observation $x_i$ means:

*Step 1*. Determine its cluster of membership;
*Step 2*. Establish whether it is a good or a bad observation in that cluster.

Let $\hat{\boldsymbol{z}}_i$ and $\hat{\boldsymbol{u}}_i$ denote, respectively, the expected values of $\boldsymbol{z}_i$ and $\boldsymbol{u}_i$ arising from the EM algorithm, i.e. $\hat{z}_{ij}$ and $\hat{v}_{ij}$ are the values of $z_{ij}^{(r)}$ and $u_{ij}^{(r)}$ at convergence. To evaluate the cluster membership of $x_i$, we use the MAP classification, i.e. MAP($\hat{z}_{ij}$). We then consider $\hat{u}_{ih}$, where $h$ is selected such that MAP($\hat{z}_{ih}$) $= 1$, and $x_i$ is considered good if $\hat{u}_{ih} > 0.5$ and $x_i$ is considered bad otherwise. The resulting information can be used to eliminate the bad points if such an outcome is desired [10]. The remaining data may then be treated as effectively being distributed according to a mixture of unimodal gamma distributions, and the clustering results can be reported as usual. Therefore, our model allows for automatic detection of bad points in the same natural way as observations are typically assigned to the groups in the finite mixture models context, i.e. based on the posterior probabilities of being good or bad points.

## 6. Numerical examples for insurance data fitting

In this section, we will evaluate the performance and various aspects of our model through artificial and real data. While the artificial data of Section 6.1 are considered to illustrate the behavior of our model in the presence of bad points, and its capability to automatically detect them, the real datasets cover different situations which may arise dealing with the empirical distribution of real-world insurance and economic data: leptokurtosis (Section 6.2), a right-heavy tail (Section 6.3), and bimodality (Section 6.4).

The proposed model is compared with several standard distributions used in the actuarial and economic literature. The analysis is conducted in R [55]. Our model has been fitted using the EM algorithm, as described in Section 4, and convergence of the algorithm is evaluated via the Aitken acceleration criterion [2,46]. The R code used to fit our model is available upon request from the authors.

In the EM algorithm, the relevance of the initialization, here the quantities $\boldsymbol{z}_i^{(0)}$, $i = 1, \ldots, n$, is well documented (see, e.g. [12,42]). For each of the applications proposed, we tried 20 different initializations and we kept the one that maximized the observed-data log-likelihood; see [7,23]. Of these 20 initializations, 18 were selected randomly, 1 using $k$-means (as implemented by the `kmeans()` function of the **stats** package), and 1 with a Gaussian mixture fitted on the log-data (as implemented by the `Mclust()` function of the **mclust** package [32]).

To compare models with the same number of parameters, in terms of goodness-of-fit, we use the log-likelihood. Comparison of models with a differing number of parameters is accomplished, as usual, via the Akaike information criterion (AIC; [4]) and the Bayesian information criterion (BIC; [58]) that, in our formulation, need to be maximized. These criteria are also used to select the number of components when mixture models are considered.

### 6.1. Artificial data

This section is based on an artificial dataset, of size $n = 500$, randomly generated by a mixture of two unimodal gamma distributions with parameters $\boldsymbol{\pi} = (0.5, 0.5)'$, $\boldsymbol{\lambda} = (1, 3)'$, and $\boldsymbol{\nu} = (0.15, 0.1)'$. The histogram of the generated data, with the true underlying density being superimposed, is displayed in Figure 2.

To evaluate the behavior of mixtures of contaminated gamma distributions in the presence of bad points, we define 45 'perturbed' datasets by adding a single outlier on the right tail. The considered outliers are equi-spaced values ranging from 6 to 50. On each of the 45 'perturbed' datasets, we fit mixtures of contaminated gamma distributions with $k \in \{1, 2, 3\}$.

In all of the considered cases, the true number of groups ($k = 2$) is always selected. Interestingly, the fitted model always detects the perturbed value as a bad point regardless of its magnitude, and this is always the unique point detected as bad. It is also interesting to note that the estimated value of $\eta_j$ (in the group containing the outlier) increases almost linearly as the value of this point further departs from the bulk of its group of membership (cf. Figure 3(a)). We also report the estimated posterior probability $\hat{u}_{501,j}$, at convergence of the EM algorithm, for the added observation to be a good point (see Equation (16) and refer to Figure 3(b)); as we can see, the farther the perturbed value is from its group, the

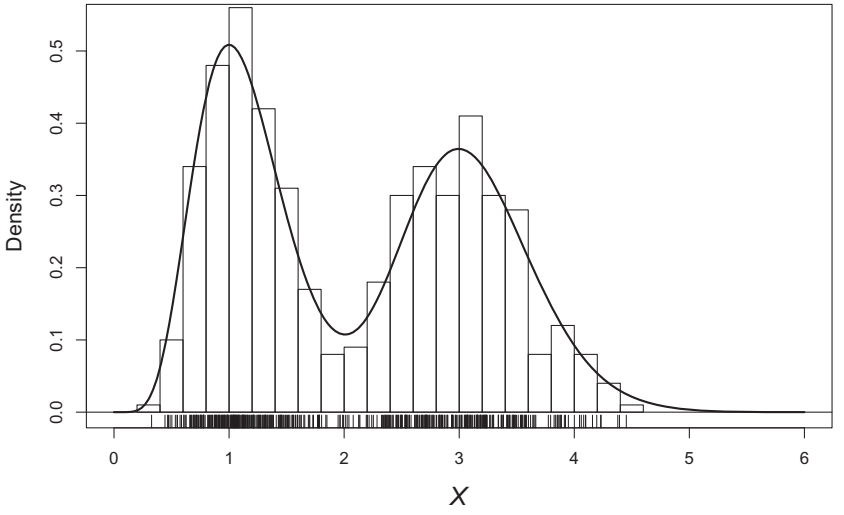


**Figure 2.** Histogram of the generated data. The underlying true mixture of two unimodal gamma densities is also superimposed.

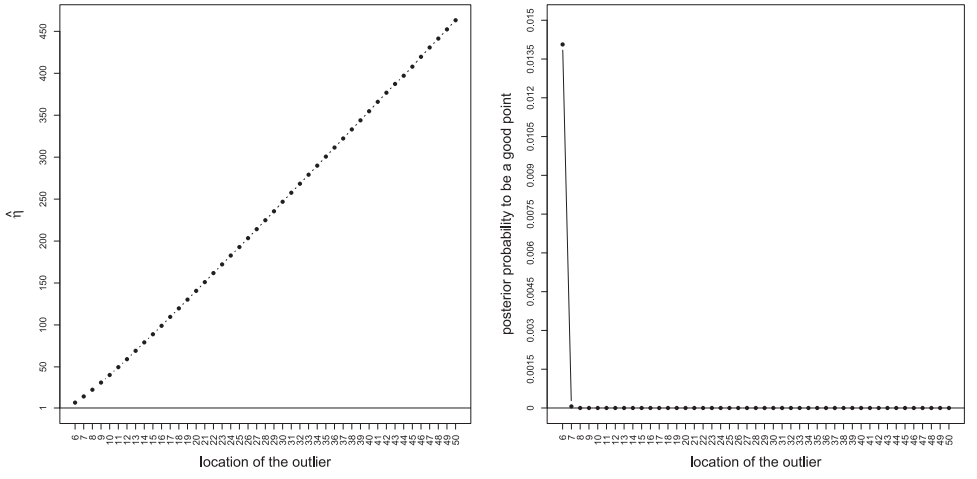


**Figure 3.** Estimated value of the inflation parameter $\eta_j$ (in the group containing the outlier) as a function of the added outlier (on the left) and posterior probability $\hat{u}_{501,j}$ to be a good point for the added outlier (in the group containing the outlier). (a) $\hat{\eta}_j$ in the group containing the outlier and (b) $\hat{u}_{501,j}$ in the group containing the outlier.

lower is its probability to be a good point, although this probability is practically null from the second case onwards (see Figure 3(b)). Such a low probability is also related to the down-weighting of this bad point in the estimation of the model parameters, and this is an important aspect for their robustness.

### 6.2. Secura Belgian Re data

The Secura Belgian Re dataset [8] – available in the R package **CASdatasets** [29] – contains $n = 371$ automobile claims from 1988 till 2001 gathered from several European insurance companies, which are at least as large as 1,200,000 euro. These data were corrected among others for inflation to reflect 2002 euros. The histogram of the claim sizes (in thousands) is depicted in Figure 4(a), while the Q–Q normal plot is given in Figure 4(b). From the Q–Q plot, a point of inflection with different slopes to the left and the right can be detected. Summary statistics are provided in Table 1.

Log-likelihood values, model selection criteria and the Kolmogorov–Smirnov (KS) test statistics are provided in Table 2, to appreciate the appropriateness of our proposal compared to the most used distributions in modeling claim sizes. Our model is fitted for a number $k$ of mixture components ranging from 1 to 4. The mixture of more than one contaminated gamma distributions performs much better (in terms of likelihood and KS statistics) than the most used distributions, e.g. the Gaussian, the skew-$t$, and more, and slightly better than very flexible ones as the Box–Cox-$t$ and the generalized inverse Gaussian, that have the drawback of properly interpreting the parameters. According to the AIC, the mixture of three contaminated gamma distributions is preferred, while the Box–Cox-$t$ model is the best one in terms of BIC.

ML estimates of model parameters for the mixture of 3 contaminated gamma distributions, along with parametrically bootstrapped standard errors over 100 replications (see, e.g. [34,50]), are provided in Table 3. The three components are reasonably well separated (see the estimates of $\lambda_1$, $\lambda_2$, and $\lambda_3$), and atypical observations arise in the right tail only (third component). The goodness of the mixture of three contaminated gamma distributions in fitting the data is also confirmed by looking at the estimated quantiles versus the historical ones (see Figure 5). It resembles almost perfectly all the quantiles of the distribution, outperforming all its competitors. On the contrary, we can note how the Box–Cox-$t$

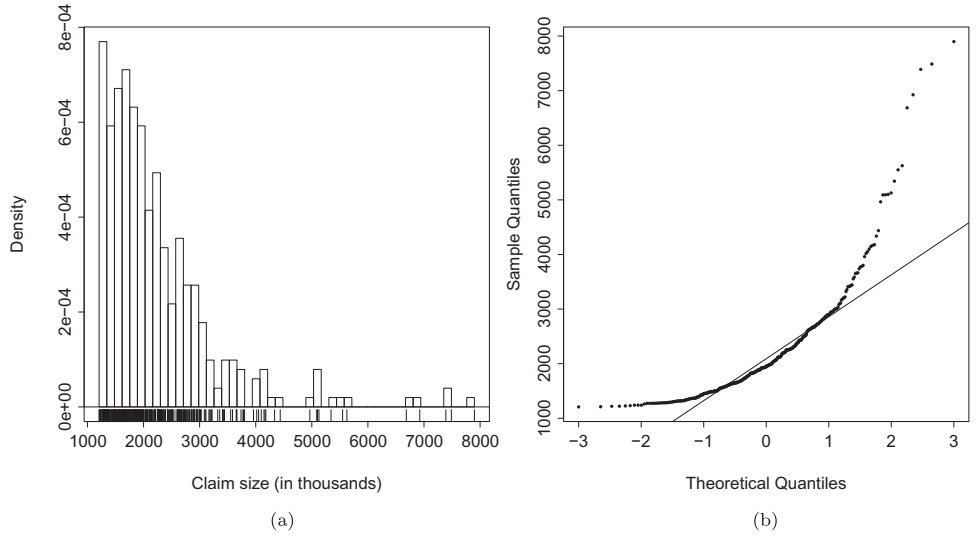


**Figure 4.** Secura Belgian Re: histogram (a) and quantile–quantile normal plot (b).

**Table 1.** Secura Belgian Re: descriptive statistics.

| Number of observations | 371 |
|---|---|
| Mean | 2230.667 |
| St.Dev | 1011.218 |
| Skewness | 2.421 |
| Kurtosis | 8.115 |
| Minimum | 1208.123 |
| Maximum | 7898.639 |
| Value at risk (95%) | 4074.796 |
| Jarque–Bera (test statistic) | 1398.800 |
| Jarque–Bera (*p*-value) | $< 0.001$ |

**Table 2.** Secura Belgian Re: goodness-of-fit measures and model selection.

| | Log-likelihood | AIC | BIC | KS statistic (*p*-value) |
|---|---|---|---|---|
| Contaminated gamma mixture ($k = 1$) | −3009.88 | −6027.76 | −6043.43 | 0.123 (0.000) |
| Contaminated gamma mixture ($k = 2$) | −2956.33 | −5930.65 | −5965.90 | 0.040 (**0.601**) |
| Contaminated gamma mixture ($k = 3$) | −2945.24 | **−5918.47** | −5973.30 | 0.026 (**0.962**) |
| Contaminated gamma mixture ($k = 4$) | −2941.83 | −5921.66 | −5996.07 | 0.022 (**0.992**) |
| Gaussian | −3092.84 | −6189.68 | −6197.52 | 0.156 (0.000) |
| Gamma | −3011.16 | −6026.31 | −6034.15 | 0.098 (0.001) |
| Log-Normal | −2984.88 | −5973.77 | −5981.60 | 0.076 (0.028) |
| Inverse Gaussian | −2985.53 | −5975.06 | −5982.89 | 0.081 (0.015) |
| Skew-*t* | −2984.48 | −5976.96 | −5992.62 | 0.074 (0.033) |
| Generalized inverse Gaussian | −2968.28 | −5942.55 | −5954.29 | 0.055 (**0.215**) |
| Box–Cox-*t* | −2956.46 | −5920.92 | **−5936.59** | 0.061 (**0.128**) |

Note: Bold font highlights the best model according to the AIC and BIC, as well as *p*-values greater than 0.05 from the KS test.

**Table 3.** Secura Belgian Re: parameter estimates and bootstrapped standard errors (in brackets) computed over 100 replications.

| $j$ | $\pi_j$ | $\lambda_j$ | $\nu_j$ | $\alpha_j$ | $\eta_j$ |
|---|---|---|---|---|---|
| 1 | 0.14 | 1338.47 | 5.66 | 0.99 | 1.01 |
| | (0.05) | (36.58) | (4.74) | (0.20) | (8.80) |
| 2 | 0.47 | 1766.18 | 43.50 | 0.99 | 1.01 |
| | (0.11) | (94.55) | (34.18) | (0.22) | (1.69) |
| 3 | 0.39 | 2635.53 | 48.24 | 0.54 | 14.26 |
| | (0.10) | (514.57) | (109.45) | (0.22) | (7.71) |

model, selected by the BIC, tends to underestimates the central quantiles, i.e. those related to probabilities approximately into the interval (0.35, 0.82), and to overestimates the quantiles on the right tail of the distribution, i.e. those related to probabilities greater than about 0.82.

### 6.3. Rent data in Munich

In this section, we would like to analyze data that at first glance may look unimodal and, thus, may be modeled by avoiding any mixture distribution. With this aim, we consider data coming from the rent survey for Munich in the year 1999 [31], available in the R package **gamlss.data**. We use the net rent – the monthly rental price, which remains after having subtracted all running costs and incidentals – per square meter as the analyzed variable. Summary statistics are provided in Table 4. Values are observed on a wide range, are

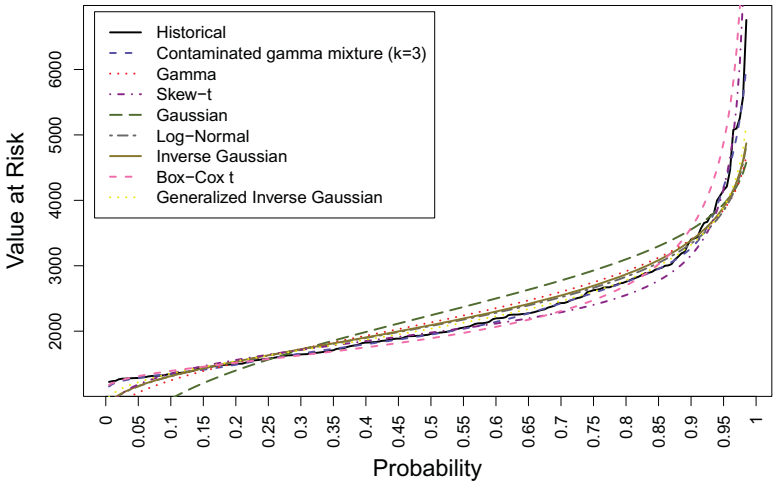


**Figure 5.** Secura Belgian Re: model fitting in terms of VaR.

**Table 4.** Rent data in Munich: descriptive statistics.

| Number of observations | 3082 |
|---|---|
| Mean | 7.111 |
| St.Dev | 2.436 |
| Skewness | 0.299 |
| Kurtosis | −0.127 |
| Minimum | 0.416 |
| Maximum | 17.722 |
| Value at risk (95%) | 11.222 |
| Jarque–Bera (test statistic) | 47.830 |
| Jarque–Bera (*p*-value) | $< 0.001$ |

almost symmetric, such that the historical VaR at the 95% is 11.222 euros and the maximum is not much greater. The histogram and the normal QQ-plot of the net rent per square meter are shown in Figure 6. The normal QQ-plot shows that the Gaussian distribution can be considered as a candidate distribution, though some problems on the left tail appear.

In the spirit of Eling [30], it is reasonable to compare the mixture of contaminated gamma distributions with more parsimonious and simple distributions to investigate the gain in model fitting, measured by the log-likelihood, obtained by our proposal. Thus, we consider a wide range of candidate distributions, with a varying number of parameters, and in Table 5 we report their log-likelihood values, along with the AIC and BIC values; results from the KS goodness-of-fit test are also displayed, since a good approximation to the data is crucial to get reliable risk measures. As before, our model is fitted for $k = 1,2,3,4$. From the last column, we note how only the mixtures of more than one contaminated gamma distribution have *p*-values greater than 5% significance level, which is the level mostly used in practical applications. This highlights the need for more flexible models for these data. According to the considered model selection criteria, the contaminated gamma mixture with $k = 2$ components is preferred according to the AIC, while the Box–Cox-*t* model is the best one for the BIC.

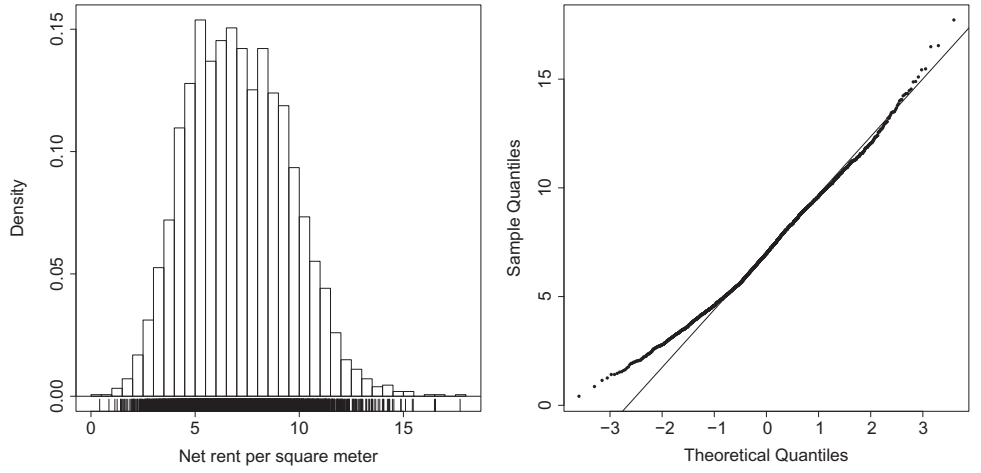


**Figure 6.** Rent data in Munich: histogram (a) and quantile–quantile normal plot (b).

**Table 5.** Rent data in Munich: goodness-of-fit measures and model selection.

| | Log-likelihood | AIC | BIC | KS statistic (*p*-value) |
|---|---|---|---|---|
| Contaminated gamma mixture ($k = 1$) | −7119.81 | −14247.63 | −14271.76 | 0.041 (0.000) |
| Contaminated gamma mixture ($k = 2$) | −7066.58 | **−14151.15** | −14205.45 | 0.008 (**0.995**) |
| Contaminated gamma mixture ($k = 3$) | −7063.85 | −14155.71 | −14240.17 | 0.007 (**0.999**) |
| Contaminated gamma mixture ($k = 4$) | −7063.84 | −14165.68 | −14280.31 | 0.007 (**0.999**) |
| Gaussian | −7116.76 | −14237.52 | −14249.59 | 0.040 (0.000) |
| Gamma | −7119.81 | −14243.61 | −14255.68 | 0.040 (0.000) |
| Log-Normal | −7231.67 | −14467.35 | −14479.41 | 0.053 (0.000) |
| Inverse Gaussian | −7279.68 | −14563.35 | −14575.42 | 0.062 (0.000) |
| Skew-*t* | −7109.86 | −14227.71 | −14251.84 | 0.040 (0.000) |
| Generalized inverse Gaussian | −7119.81 | −14245.62 | −14263.72 | 0.040 (0.000) |
| Box–Cox-*t* | −7083.56 | −14175.12 | **−14199.25** | 0.025 (0.046) |

Note: Bold font highlights the best model according to the AIC and BIC, as well as *p*-values greater than 0.05 from the KS test.

**Table 6.** Rent data in Munich: parameter estimates and bootstrapped standard errors (in brackets) computed over 100 replications.

| $j$ | $\pi_j$ | $\lambda_j$ | $\nu_j$ | $\alpha_j$ | $\eta_j$ |
|---|---|---|---|---|---|
| 1 | 0.57 | 5.14 | 0.29 | 0.50 | 4.16 |
| | (0.07) | (0.18) | (0.07) | (0.07) | (0.90) |
| 2 | 0.43 | 8.44 | 0.32 | 0.99 | 1.01 |
| | (0.07) | (0.22) | (0.06) | (0.22) | (0.77) |

ML estimates of model parameters for the mixture of 2 contaminated gamma distributions, along with parametrically bootstrapped standard errors over 100 replications are provided in Table 6.

The two components have separated modes $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \hat{\lambda}_2)' = (5.14, 8.44)'$ and similar concentration parameters $\hat{\boldsymbol{\nu}} = (\hat{\nu}_1, \hat{\nu}_2)' = (0.29, 0.32)'$. The major difference lies in the presence of outliers in one of the two components, with $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2)' = (0.50, 0.99)'$ and $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \hat{\eta}_2)' = (4.16, 1.01)'$.

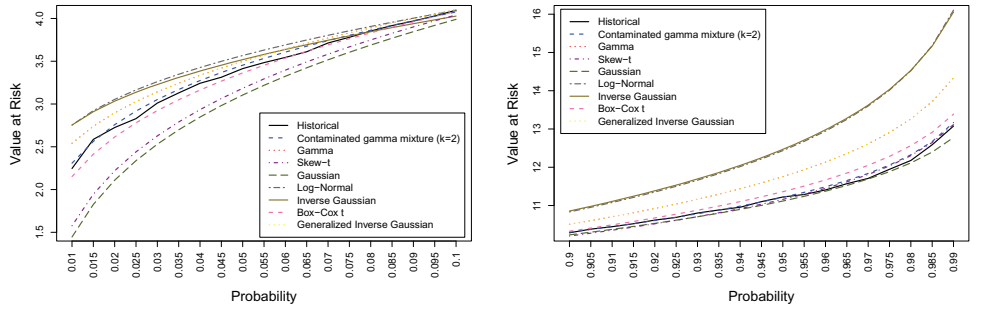


**Figure 7.** Rent data in Munich: model fitting in terms of VaR: (a) left tail and (b) right tail.

As said above, based on the KS test, none of the unimodal distributions fits well the data, neither the very flexible four-parameter Box–Cox-*t* distribution. Such a result is reflected in the analysis of the VaR under different quantiles; results are displayed in Figure 7. It is clear that some distributions are neither adequate to fit the data nor to capture the right-tail behavior of the distribution. Others, like the Gaussian and the skew-*t*, perform very well and provide close results to the historical values (see Figure 7(a)). This is not surprising as we have already noticed that the QQ-plot shows an almost Gaussian behavior, in the right tail in particular. The contaminated gamma mixture is the best model as it resembles the right behavior properly, and it has also a good performance on the left tail of the distribution. For the competing distributions, we also check for the behavior for low quantiles. The contaminated gamma mixture still captures the tail behavior, whilst both the Gaussian and the skew-*t* distributions dramatically fail to estimate low quantiles (see Figure 7(b)).

### 6.4. Film revenues

The objective of the distribution system of movies is to maximize revenues. In the North American market during the 1990s, films were released simultaneously to as many screens as distributors and exhibitors thought desirable. The dataset, available in the R package **gamlss.data**, is derived from industry standard data sourced by Nielsen EDI for the North American market annually for 13 years from 1988 to 1999, and also analyzed in [62]. All revenues are expressed in 1987 US dollars. During this period, 4031 films were released with box office opening log-revenues ranging from 4.212 to over 18.068 (see Table 7 for all the descriptive statistics).

The distribution of the data is clearly bimodal, and far from being Gaussian (see Figure 8). The superiority of the contaminated gamma mixture with respect to unimodal distributions is obvious, and this does not deserve any further investigation. We provide evidence that our approach is competitive with respect to other 2-component mixtures, widely known in the clustering literature (see Table 8).

Differences in terms of AIC and BIC are relatively small in most cases, and though the contaminated gamma mixture is preferred, other distributions are sound for the analysis of these data. To appreciate the fit of the contaminated gamma mixture, refer to Figure 9. ML estimates of the parameters are reported in Table 9.

**Table 7.** Film revenues on the log-scale: descriptive statistics.

| Number of observations | 4031 |
|---|---|
| Mean | 11.783 |
| St.Dev | 3.068 |
| Skewness | 0.037 |
| Kurtosis | −1.329 |
| Minimum | 4.212 |
| Maximum | 18.068 |
| Value at risk (95%) | 16.210 |
| Jarque–Bera (test statistic) | 297.310 |
| Jarque–Bera ($p$-value) | $< 0.001$ |

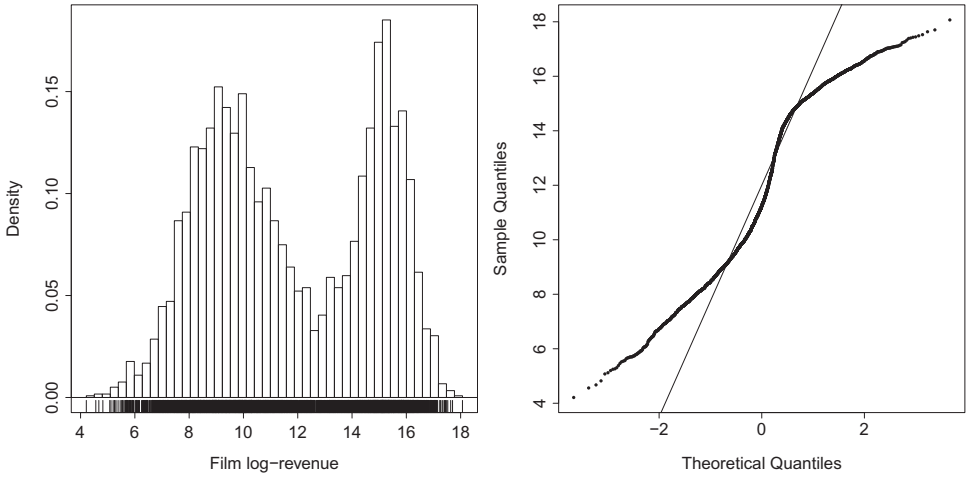


**Figure 8.** Film revenues: histogram (a) and quantile–quantile normal plot (b).

**Table 8.** Film revenues: goodness-of-fit measures, and model selection, for various mixtures with $k = 2$ components.

| Mixture component | Log-likelihood | AIC | BIC |
|---|---|---|---|
| Contaminated gamma | −9392.65 | **−18,803.31** | **−18,860.02** |
| Gaussian | −9429.05 | −18,868.10 | −18,899.60 |
| Gamma | −9437.05 | −18,884.10 | −18,915.60 |
| Log-Normal | −9494.35 | −18,998.70 | −19,030.20 |
| Inverse Gaussian | −9417.35 | −18,844.70 | −18,876.20 |
| Skew-$t$ | −9413.00 | −18,844.00 | −18,900.70 |

Note: Bold font highlights the best model according to the AIC and BIC.

The two modes are $\hat{\boldsymbol{\lambda}} = (9.43, 15.17)'$ and the concentration parameters are $\hat{\boldsymbol{\nu}} = (0.25, 0.05)'$. The major difference lies in the presence of few outliers in the first component, with $\hat{\boldsymbol{\alpha}} = (0.50, 0.99)'$ and $\hat{\boldsymbol{\eta}} = (1.90, 1.01)'$. Since additional information about this dataset is available, we can easily interpret the two clusters. Over 1764 films distributed by independent movie producers, 1651 are clustered into the first component, the one with lower revenues. This is not surprising, provided that the movie market generally produces important revenues only for productions of the major film studios, known as majors.

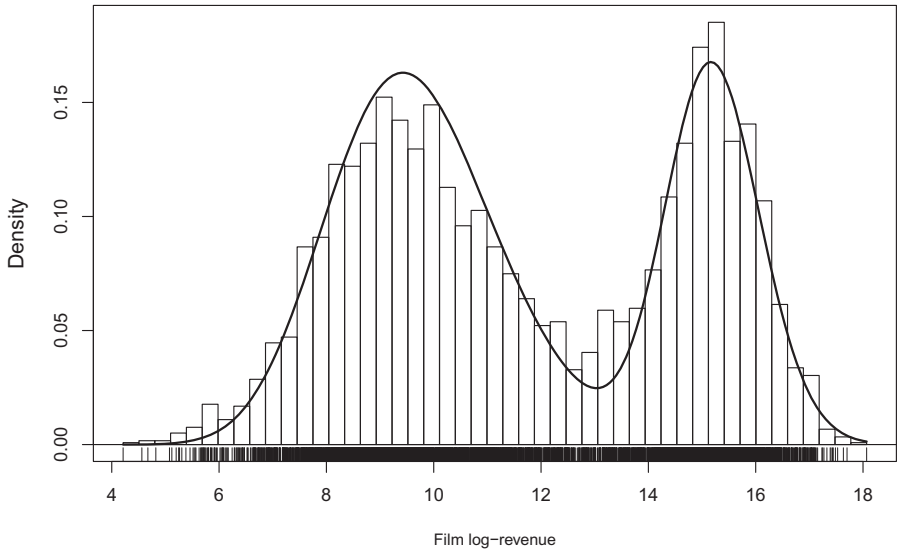


**Figure 9.** Film revenues: histogram with superimposed fitted mixture of two contaminated gamma densities.

**Table 9.** Film revenues: parameter estimates and bootstrapped standard errors (in brackets) computed over 100 replications.

| $j$ | $\pi_j$ | $\lambda_j$ | $\nu_j$ | $\alpha_j$ | $\eta_j$ |
|---|---|---|---|---|---|
| 1 | 0.63 | 9.43 | 0.25 | 0.50 | 1.90 |
| | (0.01) | (0.04) | (0.04) | (0.16) | (0.41) |
| 2 | 0.37 | 15.17 | 0.05 | 0.99 | 1.01 |
| | (0.01) | (0.03) | (0.01) | (0.22) | (0.72) |

Indeed, most films produced by the majors are clustered in the second mixture component. Nevertheless, not all movies from the majors produce high revenues, as around 37% of them belong to the first mixture component.

## 7. Conclusions

In this paper, the problem of fitting insurance and economic data was addressed by considering a finite mixture approach. We introduced finite mixtures of contaminated gamma densities able to properly account for several features typical of this type of data, such as a positive support, skewness, leptokurtosis, and multi-modality. One benefit of using this family of mixtures is that it includes the unimodal gamma distribution, mixtures of Erlang densities [45], and mixtures of unimodal gamma distributions [6] as special cases. The properties of the model are given, with accent on parameters estimation. The use of latent (or unobservable) components makes the model generic enough to handle a variety of complex real-world data, while the relatively simple prior dependence structure still allows for the use of efficient computational procedures. The obtained results, based on artificial and real data, offer a clear and coherent framework for the analysis of the typical insurance

and economic data features and confirm that the finite mixture approach is well suited to depict the process that generates the data.

Compared with the results of the most used parametric models, we found that the fit of our model is more accurate. Not only it does the flexibility required to model the data, but it also provides a better performance in terms of accuracy in evaluating measures related to the fitted distribution, as the Value-at-Risk.

We implement the fitting procedure in R and the code is available upon request from the first author. Future research may explore incorporating regressors in the mixture model, introducing the flexibility of our approach in a regression context. Furthermore, the proposed approach could be made more parsimonious by imposing convenient constraints on the parameters.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

The work of Antonello Maruotti is developed under the PRIN2015 supported-project 'Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)', funded by Ministero dell'Istruzione, dell'Universitá e della Ricerca MIUR (Italian Ministry of Education, University and Scientific Research).

## ORCID

*Antonio Punzo* http://orcid.org/0000-0001-7742-1821
*Angelo Mazza* http://orcid.org/0000-0002-5225-7452
*Antonello Maruotti* http://orcid.org/0000-0001-8377-9950

## References

[1] C.C. Aggarwal, *Outlier Analysis*, Springer, New York, 2013.
[2] A.C. Aitken, *A series formula for the roots of algebraic and transcendental equations*, Proc. Roy. Soc. Edinburgh 45 (1926), pp. 14–22.
[3] M. Aitkin and G.T. Wilson, *Mixture models, outliers, and the EM algorithm*, Technometrics 22 (1980), pp. 325–331.
[4] H. Akaike, *A new look at the statistical model identification*, IEEE. Trans. Automat. Contr. 19 (1974), pp. 716–723.
[5] A. Azzalini, T. Del Cappello, and S. Kotz, *Log-skew-normal and log-skew-t distributions as models for family income data*, J. Income Distrib. 11 (2002), pp. 12–20.
[6] L. Bagnato and A. Punzo, *Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm*, Comput. Stat. 28 (2013), pp. 1571–1597.
[7] L. Bagnato, A. Punzo, and M.G. Zoia, *The multivariate leptokurtic-normal distribution and its application in model-based clustering*, Can. J. Statist. 45 (2017), pp. 95–119.
[8] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers, *Statistics of Extremes: Theory and Applications*, Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ, 2004.
[9] L.-G. Benckert, *The lognormal model for the distribution of one claim*, ASTIN Bull. 2 (1962), pp. 9–23.
[10] M. Berkane and P.M. Bentler, *Estimation of contamination parameters and identification of outliers in multivariate data*, Sociol. Methods. Res. 17 (1988), pp. 55–64.
[11] M. Bernardi, A. Maruotti, and L. Petrella, *Skew mixture models for loss distributions: A Bayesian approach*, Insurance Math. Econom. 51 (2012), pp. 617–623.

[12] C. Biernacki, G. Celeux, and G. Govaert, *Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models*, Comput. Stat. Data. Anal. 41 (2003), pp. 561–575.
[13] T. Bodnar and A.K. Gupta, *Robustness of the inference procedures for the global minimum variance portfolio weights in a skew-normal model*, Eur. J. Finance 21 (2015), pp. 1176–1194.
[14] J. Bulla, *Hidden Markov models with t components. Increased persistence and other aspects*, Quant. Finance 11 (2011), pp. 459–475.
[15] K. Burnecki, J. Janczura, and R. Weron, *Building loss models*, in Statistical Tools for Finance and Insurance, 2nd ed., P. Cizek, W. Härdle, and R. Weron, eds., Springer-Verlag, Berlin, 2011, pp. 293–328.
[16] E. Cantoni and E. Ronchetti, *A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures*, J. Health. Econ. 25 (2006), pp. 198–213.
[17] S.X. Chen, *Probability density function estimation using gamma kernels*, Ann. Inst. Stat. Math. 52 (2000), pp. 471–480.
[18] S.T.B. Choy and C.M. Chan, *Scale mixtures distributions in insurance applications*, ASTIN Bull. 33 (2003), pp. 93–104.
[19] S.B. Choy, J.S. Chan, and U.E. Makov, *Robust bayesian analysis of loss reserving data using scale mixtures distributions*, J. Appl. Stat. 43 (2016), pp. 396–411.
[20] P. Coretto and C. Hennig, *Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering*, preprint (2016). Available at https://arxiv.org/abs/1309.6895.
[21] F.A. Cowell and M.-P. Victoria-Feser, *Robustness properties of inequality measures*, Econometrica 64 (1996), pp. 77–101.
[22] S.L. Crawford, *An application of the laplace method to finite mixture distributions*, J. Am. Stat. Assoc. 89 (1994), pp. 259–267.
[23] U.J. Dang, A. Punzo, P.D. McNicholas, S. Ingrassia, and R.P. Browne, *Multivariate response and parsimony for Gaussian cluster-weighted models*, J. Classif. 34 (2017), pp. 4–34.
[24] L. Davies and U. Gather, *The identification of multiple outliers*, J. Am. Stat. Assoc. 88 (1993), pp. 782–792.
[25] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Ser. B 39 (1977), pp. 1–38.
[26] D.K. Dey, L. Kuo, and S.K. Sahu, *A bayesian predictive approach to determining the number of components in a mixture distribution*, Stat. Comput. 5 (1995), pp. 297–305.
[27] M. Di Zio, U. Guarnera, and R. Rocci, *A mixture of mixture models for a classification problem: The unity measure error*, Comput. Stat. Data. Anal. 51 (2007), pp. 2573–2585.
[28] D.L. Donoho and P.J. Huber, *The notion of breakdown point*, in *A Festschrift For Erich L. Lehmann*, P.J. Bickel, K. Doksum, and J.L. Hodges, eds., Wadsworth Statistics/Probability Series, Wadsworth, Belmont, CA, 1983, pp. 157–184.
[29] C. Dutang and A. Charpentier, **CASdatasets**: Insurance datasets (Official website). Version 1.0-6 (2016-05-28), 2016. Available at http://cas.uqam.ca/.
[30] M. Eling, *Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models?*, Insurance Math. Econom. 51 (2012), pp. 239–248.
[31] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx, *Regression: Models, methods and applications*, Springer, Berlin, 2013.
[32] C. Fraley, A.E. Raftery, T.B. Murphy, and L. Scrucca, **mclust** Version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation, 2012.
[33] O.L. Gebizlioglu, ŞenoğluB., and Y.M. Kantar, *Comparison of certain value-at-risk estimation methods for the two-parameter weibull loss distribution*, J. Comput. Appl. Math. 235 (2011), pp. 3304–3314.
[34] W. Guan, *From the help desk: Bootstrapped standard errors*, Stata J. 3 (2003), pp. 71–80.
[35] D. Hawkins, *Identification of Outliers*, Monographs on Statistics and Applied Probability, Springer, Netherlands, 2013.
[36] C. Hennig, *Fixed point clusters for linear regression: Computation and comparison*, J. Classif. 19 (2002), pp. 249–276.

[37] C. Hennig, *Breakdown points for maximum likelihood estimators of location-scale mixtures*, Ann. Statist. 32 (2004), pp. 1313–1340.
[38] C. Hennig and P. Coretto, *The noise component in model-based cluster analysis*, in *Data Analysis, Machine Learning and Applications*, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, eds., Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, Heidelberg, 2008, pp. 127–138.
[39] C. Hennig and T.F. Liao, *How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification*, J. R. Stat. Soc. Ser. C 62 (2013), pp. 309–369. Available at http://dx.doi.org/10.1111/j.1467-9876.2012.01066.x.
[40] M. Ibragimov, R. Ibragimov, and J. Walden, *Heavy-Tailed Distributions and Robustness in Economics and Finance*, Lecture Notes in Statistics, vol. 214, Springer, New York, 2015. Available at https://books.google.it/books?id = kG6nCQAAQBAJ.
[41] A.J. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer, New York, 2008.
[42] D. Karlis and E. Xekalaki, *Choosing initial values for the EM algorithm for finite mixtures*, Comput. Stat. Data. Anal. 41 (2003), pp. 577–590.
[43] S.A. Klugman, H.H. Panjer, and G.E. Willmot, *Loss Models: From Data to Decisions*, Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ, 2012.
[44] M.N. Lane, *Pricing risk transfer transactions*, Astin Bull. 30 (2000), pp. 259–293.
[45] S.C.K. Lee and X.S. Lin, *Modeling and evaluating insurance losses via mixtures of Erlang distributions*, N. Am. Actuar. J. 14 (2010), pp. 107–130.
[46] B.G. Lindsay, *Mixture Models: Theory, Geometry and Applications*, Vol. 5. *NSF-CBMS Regional Conference Series in Probability and Statistics*, Institute of Mathematical Statistics, Hayward, California, 1995.
[47] A. Maruotti and A. Punzo, *Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers*, Comput. Stat. Data. Anal. 113 (2017), pp. 475–496.
[48] A. Maruotti, V. Raponi, and F. Lagona, *Handling endogeneity and nonnegativity in correlated random effects models: Evidence from ambulatory expenditure*, Biom. J. 58 (2016), pp. 280–302.
[49] A. Punzo, L. Bagnato, and A. Maruotti, Compound unimodal distributions for insurance losses, Insurance Math. Econom. (2017). doi:10.1016/j.insmatheco.2017.10.007.
[50] A. Punzo, R.P. Browne, and P.D. McNicholas, *Hypothesis testing for mixture model selection*, J. Stat. Comput. Simul. 86 (2016), pp. 2797–2818.
[51] A. Punzo and A. Maruotti, *Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model*, J. Comput. Graph. Stat. 25 (2016), pp. 1097–1116.
[52] A. Punzo, A. Mazza, and P.D. McNicholas, ***ContaminatedMixt**: An* R *package for fitting parsimonious mixtures of multivariate contaminated normal distributions*, J. Stat. Softw. (2018), pp. 1–25.
[53] A. Punzo and P.D. McNicholas, *Parsimonious mixtures of multivariate contaminated normal distributions*, Biom. J. 58 (2016), pp. 1506–1537.
[54] A. Punzo and P.D. McNicholas, *Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model*, J. Classif. 34 (2017), pp. 249–293.
[55] R Core Team, R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. Available at https://www.R-project.org/.
[56] G. Ritter, *Robust Cluster Analysis and Variable Selection*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, vol. 137, CRC Press, Boca Raton, 2015.
[57] A.B.Z. Salem and T.D. Mount, *A convenient descriptive model of income distribution: The gamma density*, Econometrica 42 (1974), pp. 1115–1127.
[58] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. 6 (1978), pp. 461–464.
[59] H. Teicher, *Identifiability of finite mixtures*, Ann. Math. Statist. 34 (1963), pp. 1265–1269.
[60] S. Venturini, F. Dominici, and G. Parmigiani, *Gamma shape mixtures for heavy-tailed distributions*, Ann. Appl. Stat. 2 (2008), pp. 756–776.
[61] R. Vernic, *Multivariate skew-normal distributions with applications in insurance*, Insurance Math. Econom. 38 (2006), pp. 413–426.

[62] V. Voudouris, R. Gilchrist, R. Rigby, J. Sedgwick, and D. Stasinopoulos, *Modelling skewness and kurtosis with the BCPE density in GAMLSS*, J. Appl. Stat. 39 (2012), pp. 1279–1293.
[63] M. Wiper, D.R. Insua, and F. Ruggeri, *Mixtures of gamma distributions with applications*, J. Comput. Graph. Stat. 10 (2001), pp. 440–454.

## Appendix. Details about the EM algorithm

The partial derivatives of $Q_{3j}(\lambda_j, \nu_j, \eta_j \mid \boldsymbol{\psi}^{(r)})$ in Equation (19), with respect to $\lambda_j$, $\nu_j$, and $\eta_j$, are given by

$$\begin{aligned}\frac{\partial Q_{3j}\left(\lambda_j, \nu_j, \eta_j \mid \boldsymbol{\psi}^{(r)}\right)}{\partial \lambda_j} &= \frac{1}{\eta_j \nu_j} \sum_{i=1}^{n} z_{ij}^{(r)} \left\{ \left(1 - u_{ij}^{(r)}\right) \left[ \ln x_i - \ln\left(\eta_j \nu_j\right) - \psi\left(1 + \frac{\lambda_j}{\eta_j \nu_j}\right) \right] \right. \\ &\quad \left. + \eta_j u_{ij}^{(r)} \left[ \ln x_i - \ln \nu_j - \psi\left(1 + \frac{\lambda_j}{\nu_j}\right) \right] \right\}, \\ \frac{\partial Q_{3j}\left(\lambda_j, \nu_j, \eta_j \mid \boldsymbol{\psi}^{(r)}\right)}{\partial \nu_j} &= -\frac{1}{\eta_j \nu_j^2} \sum_{i=1}^{n} z_{ij}^{(r)} \left\{ \left(1 - u_{ij}^{(r)}\right) \left[ \lambda_j - x_i + \lambda_j \ln x_i \right. \right. \\ &\quad \left. - \lambda_j \ln\left(\eta_j \nu_j\right) - \lambda_j \psi\left(1 + \frac{\lambda_j}{\eta_j \nu_j}\right) \right] \\ &\quad \left. + \eta_j \nu_j + \eta_j u_{ij}^{(r)} \left[ \lambda_j - x_i + \lambda_j \ln x_i - \lambda_j \ln \nu_j - \lambda_j \psi\left(1 + \frac{\lambda_j}{\nu_j}\right) \right] \right\},\end{aligned}$$

and

$$\begin{aligned}&\frac{\partial Q_{3j}\left(\lambda_j, \nu_j, \eta_j \mid \boldsymbol{\psi}^{(r)}\right)}{\partial \eta_j} \\ &= \frac{1}{\eta_j^2 \nu_j} \sum_{i=1}^{n} z_{ij}^{(r)} \left(1 - u_{ij}^{(r)}\right) \left[ x_i - \lambda_j - \lambda_j \ln x_i + \lambda_j \ln\left(\eta_j \nu_j\right) + \lambda_j \psi\left(1 + \frac{\lambda_j}{\eta_j \nu_j}\right) - \eta_j \nu_j \right],\end{aligned}$$

where $\psi(\cdot)$ is the digamma function.