Modelling insurance losses via contaminated unimodal distributions

Modellizzazione delle perdite assicurative tramite distribuzioni contaminate unimodali

Salvatore Daniele Tomarchio and Antonio Punzo

Abstract  Forecast the loss associated with a claim is crucial in insurance industry. These types of payments are generally highly positively skewed and with heavy tails, highlighting the necessity of flexible models. Contaminated models are a profitable way to accommodate situations in which some of the probability masses are shifted to the tails of the distribution, and in this work a general approach to contaminate unimodal hump-shaped distributions defined on a positive support is introduced. The proposed models are hence fitted to a real insurance loss dataset, along with several standard distributions used in the actuarial literature. Comparison between the models is made using information criteria and risk measures such as VaR and TVaR.

Abstract  Prevedere la perdita associata alle richieste di risarcimento, è fondamentale in campo assicurativo. Questo tipo di pagamenti sono, in genere, positivamente asimmetrici e a code pesanti, evidenziando quindi il bisogno di avere modelli flessibili. I modelli contaminati sono uno strumento utile per gestire situazioni nelle quali si hanno masse di probabilità spostate sulle code, ed in questo lavoro viene introdotto un approccio generale per contaminare distribuzioni unimodali definite su supporto positivo. I modelli proposti sono quindi testati su un dataset reale riguardante perdite assicurative, insieme a diverse altre distribuzioni standard usate in letteratura. I confronti tra i modelli sono fatti usando dei criteri informativi e due misure di rischio come il VaR ed il TVaR.

Key words: Insurance losses, Contaminated model, Value at Risk, Tail Value at Risk

Salvatore Daniele Tomarchio
Department of Economics and Business, University of Catania, Corso Italia, 55, 95129 Catania, Italy, e-mail: daniele.tomarchio@unict.it

Antonio Punzo
Department of Economics and Business, University of Catania, Corso Italia, 55, 95129 Catania, Italy, e-mail: antonio.punzo@unict.it
1 Introduction

It is crucial, in insurance business, to find adequate models for loss data in order to correctly compute premiums, risk measures and the required reserves. Unfortunately, this is not an easy task because of the distinctive characteristics of their distribution. As widely documented, the loss distribution is unimodal hump-shaped [4], highly positively skewed [11], and with heavy tails [1]. Some authors argued that observed losses can be described by a single probability distribution [6, 7, 10, 12]. However, some of these distributions are defined on the whole real line, causing the so-called boundary bias problem [8], while others fail to cover the behaviour of either small or high losses [9]. In particular, the losses in the upper tail, though rare in frequency, are the ones that have the most impact on the financial stability of insurance companies. Considering this, in Section 2 we propose a contaminated approach that allows to account for all the peculiarities of the loss data discussed above, with particular reference to the tails behaviour of the distribution. In detail, a 2-parameter unimodal hump-shaped model, reparameterized with respect to the mode \( \lambda \) and to another parameter \( \nu \) that is strictly related to the distribution variability, is chosen as “core distribution”. An analogous distribution, in which \( \nu \) is multiplied by another parameter \( \eta > 1 \), is chosen as “contaminant distribution”. The mix of these two distributions generates a 4-parameter contaminated model being unimodal in \( \lambda \) and giving more flexibility to the tails with respect to the core distribution. Furthermore, the proposed models allow for automatic detection of ‘bad’ losses via a simple procedure based on maximum a posteriori probabilities. According to our approach, and in the fashion of Aitkin and Wilson [2], bad losses are defined with respect to the core distribution as points producing an overall distribution (i.e. the contaminated distribution) that is too heavy-tailed in order to be modeled by the core distribution only. In other words, endowed with heavy tails our model offers the flexibility needed for achieving bad losses robustness, whereas the core distribution lacks sufficient fit. Two examples of contaminated models are examined in Section 2, and then fitted to a real insurance loss dataset, along with other well-known parametric distributions, in Section 3. Comparisons between the models are made using information criteria and risk measures, while some conclusions, as well as future possible extensions, are drawn in Section 4.

2 Contaminated models: A general framework and two specific applications

Let \( X \) be a positive random variable. Requiring that the probability density function (pdf) \( p(x) \) of \( X \) should be unimodal hump-shaped and positively skewed, a general (4-parameter) contaminated unimodal pdf for losses could be

\[
p(x; \theta) = \alpha f(x; \lambda, \nu) + (1 - \alpha) f(x; \lambda, \eta \nu), \quad x > 0,
\]  

(1)
where $\vartheta = (\lambda, \nu, \eta, \alpha)'$. In (1), $f(x; \lambda, \nu)$ and $f(x; \lambda, \eta \nu)$ are the unimodal hump-shaped densities selected as core and contaminant distributions, respectively. $\lambda > 0$ is the mode, $\nu > 0$ is a parameter that manages the concentration of $f$ around the mode, $\eta > 1$ indicates the degree of contamination, and can be interpreted as the increase in variability due to the excessively small or large losses with respect to the core distribution, whereas $\alpha \in (0, 1)$ is the weight applied to the core distribution. It should be noted that, because both distributions have their maximum in $\lambda$, even the resulting contaminated model $p(x)$ will have mode $\lambda$. Among all the existing 2-parameter distributions that can be used for $f$, unimodal gamma and log-normal will be considered.

Let $f(x; \alpha, \beta)$ be the pdf of a gamma distribution with the standard parameterization, i.e. where $\alpha > 0$ and $\beta > 0$ are the shape and scale parameters, respectively. In order to have a core distribution for losses that can be inserted in model (1), a reparameterization is needed. Setting

$$
\begin{aligned}
\alpha &= \frac{\lambda}{\nu} + 1 \\
\beta &= \nu
\end{aligned}
\Rightarrow
\begin{aligned}
\lambda &= \beta (\alpha - 1) \\
\nu &= \beta
\end{aligned},
\tag{2}
$$

we obtain

$$
f(x; \lambda, \nu) = \frac{x^{\frac{\lambda}{\nu} - 1} e^{-x}}{\frac{1}{\nu} + 1 \Gamma \left( \frac{\lambda}{\nu} + 1 \right)}, \quad x > 0,
\tag{3}
$$

with $\lambda > 0$ and $\nu > 0$. More details about this type of parameterization can be found in [8] and [3]. Ultimately, it should be clarified that only the subset of unimodal gamma densities is considered, neglecting all the unlimited reverse J-shaped ones that have a vertical asymptote in $x = 0$.

Let $f(x; \mu, \sigma)$ be the pdf of a log-normal distribution with the standard parameterization where $\mu \in \mathbb{R}$ and $\sigma > 0$. With the purpose of having a core distribution for losses that can be inserted in model (1), also in this case, a reparameterization is needed. Imposing

$$
\begin{aligned}
\mu &= \ln \lambda + \nu \\
\sigma^2 &= \nu
\end{aligned}
\Rightarrow
\begin{aligned}
\lambda &= e^{\mu - \sigma^2} \\
\nu &= \sigma^2
\end{aligned},
\tag{4}
$$

the pdf becomes

$$
f(x; \lambda, \nu) = \frac{e^{\frac{(\ln x - \ln \lambda - \nu)^2}{2\sigma^2}}}{\sqrt{2\pi \nu \lambda}}, \quad x > 0,
\tag{5}
$$

with $\lambda > 0$ and $\nu > 0$.

An interesting characteristic of model (1) is that, once $\vartheta$ is estimated, say $\hat{\vartheta}$, it is possible to determine whether a generic loss, say $x^*$, is good via the a posteriori probability

$$
P \left( x^* \text{ is good} \bigg| \hat{\vartheta} \right) = \frac{\hat{\alpha} f \left( x^*; \hat{\lambda}, \hat{\nu} \right)}{p \left( x^*; \hat{\vartheta} \right)}.
\tag{6}
$$
Specifically, $x^*$ will be considered good if $P\left(x^* \text{ is good} \mid \hat{\theta}\right) > 1/2$, while it will be considered bad otherwise.

## 3 Application to insurance loss dataset

The dataset consists of 2,387 French business interruption losses over the period 1985 to 2000. For each observation, total cost (that includes the additional expenses associated with settlement of the claim) in French francs (FF) is considered. Comparisons between distributions are presented in Table 1. AIC and BIC indicate that

<table>
<thead>
<tr>
<th>Model</th>
<th>Log-lik</th>
<th>AIC</th>
<th>BIC</th>
<th>Rank</th>
<th>LR test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cont.gamma</td>
<td>-19,983.29</td>
<td>-39,974.58</td>
<td>3</td>
<td>-39,997.69</td>
<td>3 0.000</td>
</tr>
<tr>
<td>Cont.log-normal</td>
<td>-19,842.98</td>
<td>-39,693.97</td>
<td>1</td>
<td>-39,717.08</td>
<td>1 0.000</td>
</tr>
<tr>
<td>Exponential</td>
<td>20,563.23</td>
<td>-41,128.46</td>
<td>6</td>
<td>-41,134.23</td>
<td>6 0.000</td>
</tr>
<tr>
<td>Gamma (unimodal)</td>
<td>20,563.23</td>
<td>-41,130.46</td>
<td>7</td>
<td>-41,142.01</td>
<td>7 0.000</td>
</tr>
<tr>
<td>Log-normal</td>
<td>19,893.29</td>
<td>-39,790.59</td>
<td>2</td>
<td>-39,802.14</td>
<td>2 0.000</td>
</tr>
<tr>
<td>Weibull</td>
<td>20,254.73</td>
<td>-40,513.47</td>
<td>5</td>
<td>-40,525.02</td>
<td>5 0.000</td>
</tr>
<tr>
<td>Normal</td>
<td>24,127.48</td>
<td>-48,258.95</td>
<td>12</td>
<td>-48,270.51</td>
<td>12 0.000</td>
</tr>
<tr>
<td>Cauchy</td>
<td>20,769.81</td>
<td>-41,543.62</td>
<td>8</td>
<td>-41,555.17</td>
<td>8 0.000</td>
</tr>
<tr>
<td>Logistic</td>
<td>22,261.65</td>
<td>-44,527.29</td>
<td>10</td>
<td>-44,538.85</td>
<td>10 0.000</td>
</tr>
<tr>
<td>Skew-logistic</td>
<td>21,720.53</td>
<td>-43,447.07</td>
<td>9</td>
<td>-43,464.40</td>
<td>9 0.000</td>
</tr>
<tr>
<td>Skew-normal</td>
<td>22,592.65</td>
<td>-45,191.31</td>
<td>11</td>
<td>-45,208.64</td>
<td>11 0.000</td>
</tr>
<tr>
<td>Skew-t</td>
<td>20,039.17</td>
<td>-40,086.33</td>
<td>4</td>
<td>-40,109.44</td>
<td>4 0.000</td>
</tr>
</tbody>
</table>

Table 1. French business interruption losses: log-likelihood, AIC, and BIC for the competing models, along with rankings. In the last column, $p$-values from the LR tests.

The cont.log-normal model is the best one, while the cont.gamma is ranked third. They further provide an improvement compared to their core distributions, as confirmed by the null $p$-values of the LR test. Table 2 reports the empirical and the estimated VaR and TVaR of the fitted models, at confidence levels of 95% and 99%. The ranking here is based on the absolute value of the percentage of variation with respect to the empirical risk measure considered; the lower the difference the better is the position in the ranking. A backtesting procedure is also applied to test when models provide reasonable estimates of the VaR. Analysing the VaR, at the 95% confidence level the cont.log-normal is again the best model, and it seems to be the only one to reproduce the empirical VaR, with a $p$-value very close to 1. At the 99% confidence level, the best model is the cont.gamma instead. If $p$-values are checked, both contaminated models seem able to reproduce the empirical VaR. Considering now the TVaR, at both confidence level the cont.log-normal is the best model, while the cont.gamma is the second best. Nevertheless, only the cont.log-normal assumes a very good value, considering that all the others are even further away from the true value than the preceding case.
Table 2 French business interruption losses: VaR, with its backtest, and TVaR at confidence levels of 95% and 99%.

Finally, to show how formula (6) works, the largest ten losses are considered in Table 3 and Table 4. As stated in Section 1, these losses could be considered like outliers that contaminate the core distribution, implying a heavier right tail than expected. Therefore, in a contaminated model, they should belong to the contaminant distribution, and treated like bad losses if such an outcome is desired [5].

4 Conclusions

In this paper, a general contaminated model has been introduced by mixing a core distribution with a contaminant one. By using a contamination approach, as pro-
posed here, both small and large observations can be accommodated and hence reliable statistical inference is possible also for heavy-tailed loss distributions. The main finding is that both models behave very well compared to the 12 benchmark distributions considered, both in terms of goodness of fit and in the computation of risk measures. A logical extension of this work would be to allow also for other 2-parameter unimodal hump-shaped distributions (defined on a positive support) to be used as core and contaminant distributions and to apply these models to a variety of other insurance loss datasets. In the fashion of Punzo and McNicholas [13], define mixtures of our contaminated unimodal models to be used as a powerful devise for robust clustering and density estimation of positive data.

References