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 stumento 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 λ and to another parameter ν that is strictly related to the distribution variability, is chosen as "core distribution". An analogous distribution, in which ν 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 λ 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

$$p(x;\vartheta) = \alpha f(x;\lambda,\nu) + (1-\alpha) f(x;\lambda,\eta\nu), \quad x > 0, \tag{1}$$

where $\vartheta = (\lambda, \nu, \eta, \alpha)'$. In (1), $f(x;\lambda,\nu)$ and $f(x;\lambda,\eta\nu)$ are the unimodals hump-shaped densities selected as core and contaminant distributions, respectively. $\lambda > 0$ is the mode, $\nu > 0$ is a parameter that manage 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 λ, even the resulting contaminated model $p(x)$ will have mode λ. 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

$$\begin{cases} \alpha = \dfrac{\lambda}{\nu} + 1 \\ \beta = \nu \end{cases} \quad \Rightarrow \quad \begin{cases} \lambda = \beta(\alpha - 1) \\ \nu = \beta \end{cases}, \tag{2}$$

we obtain

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

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 unimodals 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 $\mu \in$ 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

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

the pdf becomes

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

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

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

$$P\left(x^* \text{ is good} \middle| \widehat{\vartheta}\right) = \frac{\widehat{\alpha} f\left(x^*;\widehat{\lambda},\widehat{\nu}\right)}{p\left(x^*;\widehat{\vartheta}\right)}. \tag{6}$$

Specifically, x^* will be considered good if $P\left(x^* \text{ is good} \middle| \widehat{\vartheta}\right) > 1/2$, 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

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

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 is 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 able 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.

Model	VaR				Prop. Viol.		*p*-value		TVaR			
	95%	Rank	99%	Rank	95%	99%	95%	99%	95%	Rank	99%	Rank
Empirical	7,675.81		18,293.88						17,062.59		38,135.05	
Cont.gamma	9,454.65	6	18,931.39	1	0.036	0.009	0.001	0.547	15,341.69	2	24,789.96	2
Cont.log-normal	7,787.84	1	21,810.33	5	0.049	0.006	0.899	0.050	17,872.90	1	40,499.24	1
Exponential	6,074.55	4	9,338.07	9	0.071	0.037	0.000	0.000	8,100.34	9	11,365.48	9
Gamma (unimodal)	6,074.52	5	9,338.02	10	0.071	0.037	0.000	0.000	8,102.25	8	11,365.75	8
Log-normal	6,189.57	3	14,304.85	6	0.067	0.017	0.000	0.001	11,822.42	5	23,777.22	4
Weibull	6,893.35	2	12,358.04	8	0.062	0.023	0.012	0.000	10,366.06	7	16,282.58	7
Normal	11,792.92	10	15,838.83	3	0.026	0.013	0.000	0.161	14,303.56	4	17,865.08	6
Cauchy	3,429.48	11	14,992.23	4	0.132	0.017	0.000	0.002	186,392.18	12	906,557.85	12
Logistic	5,097.50	8	7,257.37	11	0.082	0.059	0.000	0.000	6,439.81	10	8,570.36	10
Skew-Logistic	4,930.70	9	7,093.86	12	0.084	0.059	0.000	0.000	6,264.82	11	8,391.85	11
Skew-Normal	12,385.83	12	16,246.27	2	0.023	0.012	0.000	0.307	14,752.98	3	18,233.47	5
Skew-*t*	5,809.08	7	12,804.28	7	0.074	0.022	0.000	0.000	11,261.57	6	24,290.48	3

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].

Loss value (x^*)	Probability that (x^*) is good
27,440.82	0.001245
31,567.43	0.000329
32,450.93	0.000247
44,088.55	0.000006
45,500.31	0.000004
46,827.85	0.000002
50,155.73	0.000001
53,357.16	0.000000
152,449.02	0.000000
168,654.35	0.000000

Table 3 French business interruption losses: *a posteriori* probability to be a good observation for the largest 10 losses based on the contaminated gamma.

Loss value (x^*)	Probability that (x^*) is good
27,440.82	0.218933
31,567.43	0.199455
32,450.93	0.195780
44,088.55	0.158422
45,500.31	0.154925
46,827.85	0.151789
50,155.73	0.144506
53,357.16	0.138185
152,449.02	0.060963
168,654.35	0.056014

Table 4 French business interruption losses: *a posteriori* probability to be a good observation for the largest 10 losses based on the contaminated log-normal.

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

1. Ahn, Soohan, Joseph HT Kim, and Vaidyanathan Ramaswami. "A new class of models for heavy tailed distributions in finance and insurance risk." Insurance: Mathematics and Economics 51.1 (2012): 43-52.
2. Aitkin, Murray, and Granville Tunnicliffe Wilson. "Mixture models, outliers, and the EM algorithm." Technometrics 22.3 (1980): 325-331.
3. Bagnato, Luca, and Antonio Punzo. "Finite mixtures of unimodal beta and gamma densities and the k -bumps algorithm." Computational Statistics 28.4 (2013): 1571-1597.
4. Bakar, SA Abu, et al. "Modeling loss data using composite models." Insurance: Mathematics and Economics 61 (2015): 146-154.
5. Berkane, Maia, and Peter M. Bentler. "Estimation of contamination parameters and identification of outliers in multivariate data." Sociological Methods & Research 17.1 (1988): 55-64.
6. Bickerstaff, David R. "Automobile Collision Deductibles and Repair Cost Groups: The Lognormal Model." PCAS LIX (1972): 68.
7. Burnecki, Krzysztof, Adam Misiorek, and Rafal Weron. "Loss distributions." Statistical Tools for Finance and Insurance. Springer, Berlin, Heidelberg, 2005. 289-317.
8. Chen, Song Xi. "Probability density function estimation using gamma kernels." Annals of the Institute of Statistical Mathematics 52.3 (2000): 471-480.
9. Cooray, Kahadawala, and Malwane MA Ananda. "Modeling actuarial data with a composite lognormal-Pareto model." Scandinavian Actuarial Journal 2005.5 (2005): 321-334.
10. Eling, Martin. "Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models?." Insurance: Mathematics and Economics 51.2 (2012): 239-248.
11. Jeon, Yongho, and Joseph HT Kim. "A gamma kernel density estimation for insurance loss data." Insurance: Mathematics and Economics 53.3 (2013): 569-579.
12. Kazemi, Ramin, and Monireh Noorizadeh. "A Comparison between Skew-logistic and Skew-normal Distributions." Matematika 31.1 (2015): 15-24.
13. Punzo, Antonio, and Paul D. McNicholas. "Parsimonious mixtures of multivariate contaminated normal distributions." Biometrical Journal 58.6 (2016): 1506-1537.