Severity modelling with external loss data

Models of operational risk, based only on internal loss data, can exhibit wide confidence intervals for capital at 99.9 percentile of the aggregate loss distribution. Consequently, banks rely on scenario and/or external loss data to improve capital estimates and to narrow confidence intervals. Mikhail Makarov and Bahram Mirzai report

In this article we consider the use of external loss data for operational risk modelling, where examples of such pools are public data sources or loss data consortia. In the following we use the notion of pooled data instead of external data to stress that a banks own data can be part of the pooled data. We focus only on severity modelling, and, in particular, we show that a severity distribution fitted to the pooled data should not be used directly to estimate a bank's tail severity if the difference between the banks in the pool is significant.

Section 1 considers two examples of pooled data to illustrate some of the phenomena that can occur when modelling a bank's tail severity using pooled data. Section 2 seeks to explain why such phenomena are observed. In Section 3 we outline methods for modelling a bank's tail severity by taking into account the difference between the banks in the pool.

1. Examples of pooled data

To illustrate some of the phenomena encountered when modelling severity using pooled data sets, we consider two examples: the first corresponding to actual loss data and the second being a constructed example.

1.1 LDCE data

Fed Reserve Banks of Boston presents results of 2004 LDCE data[1] based on the operational risk losses of 23 participating banks, and that publication's table 7a[1] provides the following summary statistics:

All groups:

No of observations	25th percentile	50th percentile	75th percentile	95th percentile
1,120	13,546	20,738	43,574	221,271

Suppose now that we would like to fit a severity distribution starting from the threshold 20,738. Although we do not have access to the individual losses in excess of the threshold, one can still apply the maximum likelihood method to the percentiles in the table to obtain an approximate severity fit to the loss data. To accomplish this, we choose to fit a Pareto distribution with the following PDF:

$$p(x) = \frac{\alpha}{R}\left(\frac{R}{x}\right)^{\alpha+1}$$

where the threshold is fixed at R=20,738. The choice of the Pareto distribution is motivated by the fact that it corresponds to a GPD distribution with the scale parameter fixed at the selected threshold. Moreover, the Pareto distribution is a more convenient choice in view of the limited data, as only one parameter needs to be estimated.

The maximum likelihood estimate results in α=0.956. As the mean of the Pareto distribution is only finite for α>1, we obtain a heavy-tailed severity with infinite mean.

The question arises whether such a heavy-tailed distribution can be a valid representation of the tail severity for an individual bank.

1.2 Light-tailed banks/heavy-tailed pool

In this example we construct a pool consisting of banks with light-tailed severity distributions, and show that under certain assumptions the resulting severity for the pooled data is a heavy-tailed distribution.

To construct the example we choose individual banks to have an exponential severity X_θ. The PDF of the exponential distribution is:

$$p(x|\theta) = \theta \exp(-\theta\, x)$$

Notice that each bank has a finite average loss of $1/\theta$. We assume the parameter θ to follow a gamma distribution with PDF:

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)$$

For simplicity we further assume that the banks in the pool have the same loss frequency.

Under these assumptions, the severity X_{pool} of the pooled data has the following PDF:

$$p_{pool}(x) = \int p(x|\theta)\,p(\theta)\,d\theta = \alpha\beta^\alpha\left(\frac{1}{\beta+x}\right)^{\alpha+1}$$

This corresponds to the PDF of a second Pareto distribution with a shape parameter α and a scale parameter β. The second Pareto distribution belongs to the class of EVT distributions and is a heavy-tailed distribution. In particular, the mean of the distribution is given by:

$$\mathrm{E}\left[X_{pool}\right] = \frac{\beta}{\alpha-1}$$

which becomes infinite for $\alpha \le 1$.

Hence we have constructed an example where the individual banks have light-tailed severity distributions with finite mean, but the corresponding pooled data has a heavy-tailed severity distribution with possibly infinite mean. The example is also interesting in view of the LDCE data, as it shows that the heavy-tailed nature of the pool may not

necessarily imply the heavy-tailed nature of the individual banks in the pool.

2. Variance analysis

Typically, models developed for pooled data sets follow the structure of example 1.2. The severities of individual banks are assumed to belong to a distribution family $p(x|\theta)$ parameterised by θ. The parameter θ is assumed not to be arbitrary but to follow a distribution that is calibrated by means of the pool data.

For such parametric models, we would like to investigate under which conditions pool severity X_{pool} is a good approximation of an individual bank's severity X_θ. To accomplish this, we split the total variance of the pooled data according to:

$$\mathrm{Var}\left[X_{pool}\right] = \mathrm{Var}_\theta\left[\mathrm{E}\left[X_\theta|\,\theta\right]\right] + \mathrm{E}_\theta\left[\mathrm{Var}\left[X_\theta|\,\theta\right]\right]$$

The first term in this equation describes the variance between the banks, whereas the second term corresponds to the average variance within banks. For illustration, if the split is applied to example 1.2, we obtain:

$$\frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} + \frac{\beta^2}{(\alpha-1)(\alpha-2)}$$

For simplicity of notation, we rewrite the variance split equation as:

$$\mathrm{Var}_{\mathrm{total}} = \mathrm{Var}_{\mathrm{between}} + \mathrm{Var}_{\mathrm{within}}$$

In this notation, the formula suggests the following intuitive result: when the difference between banks in the pool is small, ie $\mathrm{Var}_{\mathrm{between}} \approx 0$, the pooled data severity is a good approximation of the severity of individual banks:

$$\mathrm{Var}_{\mathrm{within}} \approx \mathrm{Var}_{\mathrm{total}}$$

However, if the difference between banks becomes significant, this approximation fails and other techniques are required to estimate severity distributions for individual banks.

3. Severity modelling

We consider in this section two example approaches that can be used to estimate the severity of a bank using both the bank's internal losses and pooled data: credibility analysis and Bayesian analysis.

It is worth mentioning that credibility analysis is a linearisation of the Bayesian approach, and as such it is easier to apply, but the Bayesian analysis may lead to more accurate estimations of the severity of the individual banks. In many cases, both approaches lead to the same result.[2]

3.1 Credibility analysis

Credibility analysis combines a bank's internal data with the pooled data by assigning appropriate weights. A simple non-parametric model was developed.[3]

In order to estimate the cumulative distribution function of the k-th bank in the pool, the following approximation is applied:

$$F_\theta(x) \approx F_k(x) = z_k(x)F_{emp,\,k}(x) + \left(1 - z_k(x)\right)F_{pool}(x)$$

That is, the severity F_k of the k-th bank is estimated as the weighted average of its empirical severity $F_{emp,\,k}$ and the severity F_{pool} of the pooled data.

The credibility factor used to build the average of the two distributions is estimated as follows:

$$z_k(x) = \frac{n_k}{n_k + N(x)}$$

where n_k is the number of internal losses of the k-th bank and

$$N(x) = \frac{F_{pool}(x)\,\left(1 - F_{pool}(x)\right)}{\mathrm{Var}_\theta\,\left[F_\theta(x)\,|\,\theta\right]}$$

The previous formulas have two desired properties. Firstly, the more internal losses are observed for a bank, the higher becomes the credibility factor assigned to that bank's empirical severity. Secondly, the larger the difference between the banks in the pool, the more weight is given to each bank's empirical severity.

3.2 Bayesian analysis

The Bayesian analysis provides a parametric approach to the estimation of a bank's severity. The Bayesian approach is considered here for the case when the severities of the banks belong to the exponential family.[4,5]

The general form of an exponential family PDF is given by:[6]

$$p(x|\theta) \;=\; h(x)\exp\left(\phi(\theta)T(x) - A(\theta)\right)$$

The exponential family is a convenient choice for Bayesian analysis, as a conjugate family of priors with parameters μ and υ can be constructed:

$$p(\theta|\mu,\nu) \;\propto\; \exp\left(\phi(\theta)\mu - \nu A(\theta)\right)$$

To obtain the severity of a bank, ie the corresponding value of θ, with observed losses $\{y_1, \ldots, y_n\}$, we apply the maximum likelihood method, where the likelihood function is given by:

$$f\left(\theta|\{y_1,\ldots,y_n\}\mu,\nu\right) \propto \exp\left(\phi(\theta)\left(\mu + \sum_{i=1}^{n} T(y_i)\right) - (n+\nu)A(\theta)\right)$$

The maximum of the likelihood function corresponds to the desired value of θ that combines the internal losses of a bank with pooled data.

Conclusions

For pooled data, if the contributing banks have significantly different severities, the pool severity is likely to be more heavy-tailed than the individual severities. In order to estimate the severity of a bank using both pooled data and internal loss data, one needs to scale losses using exposure information, group banks into severity clusters, and perform credibility or Bayesian analysis on cluster data.

6. Examples of distributions belonging to exponential group are normal, lognormal, gamma, loggamma, exponential and Pareto distributions

References

1. Federal Reserve System, 2005
Results of the 2004 loss data collection exercise for operational risk
May 12

2. Jewell WS, 1974
Credible means are exact Bayesian for exponential families
ASTIN Bulletin 1974: 8; 77–90

3. Jewell WS, 1974
The credible distribution
ASTIN Bulletin 7; 237–69

4. Frangos NE, Vrontos SD, 2001
Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance
ASTIN Bulletin 31 (No 1); 1–22

5. Gelman A *et al*, 2003
Bayesian Data Analysis,
2nd edition, Chapman & Hall/CRC

Mikhail Makarov and Bahram Mirzai are managing partners at EVMTech.
Email: bmirzai@evmtech.com