Topic of the session

Dealing actuarial parametric models with R statistical software

Shaiful Anuar Abu Bakar

Motivation

- Fitting insurance loss data using Pareto, Gamma and Lognormaldistributions are very common in the actuarial field.
- However, with recent development in loss modelling, more advanced models with mixture of distributions are being proposed.
- These so called generated models has shown good performances in terms of their statistical features to describe the nature of loss data.
- In this respect, some important literatures on loss modelling include the mixture model by Miljkovic and Grun (2016), the composite models by Cooray and Ananda (2005), the folded models by
- Brazauskas and Kleefeld (2011) and the arc tan models by Gomez-Deniz and Calderin-Ojeda (2015).
- Here we reintroduce the models and provide some algorithm to deal with them.
- Besides, the significance of the model in performing various task in R statistical software are presented in brief with relevant employment of real loss data.

Mixture Model

The density of the mixture model is given by

$$f(x) = \sum_{i=1}^{n} a_i f_i(x) \quad x > 0$$

where sum of all a_is equal to one.
A two component mixture model is therefore given by

$$f(x) = af_1(x) + (1-a)f_2(x) \quad x > 0.$$

It is common that a is expressed as $\frac{1}{1+\phi}$ and thus having all parameters greater than zero (provided parameters in $f_1(x)$ and $f_2(x)$ are greater than zero).

- When dealing with real Danish fire loss data, Milkjovic and Grun (2016) showed that the 2-component Burr mixture model has the best fitting according to several information criteria selection method.
- The model also has the lowest difference between the empirical and theoretical risk measures (for VaR and CTE).

Composite model

The density function of the composite model is define as follows

$$f(x) = \begin{cases} a_1 f_1^*(x), & \text{if } 0 < x < \theta, \\ a_2 f_2^*(x), & \text{if } \theta < x < \infty. \end{cases} \tag{1}$$

where

$$f_i^*(x) = \frac{f_i(x)}{\int_a^b f_i(x)dx}$$

is the truncated distribution with $a < x < b$ and $a_1 + a_2 = 1$.

The distribution function of the composite model is given by

$$F(x) = \begin{cases} a_1 F_1^*(x), & \text{if } 0 < x < \theta, \\ a_1 + a_2 F_2^*(x), & \text{if } \theta < x < \infty. \end{cases}$$

where

$$F_i^*(x) = \frac{F_i(x) - F_i(\text{domain lower limit})}{F_i(\text{domain upper limit}) - F_i(\text{domain lower limit})}$$

Quantile function of composite model is given by

$$Q(x) = \begin{cases} Q_1^*(\frac{u}{a_1}), & \text{if } 0 < u < a_1, \\ Q_2^*(\frac{u - a_1}{a_2}), & \text{if } a_1 < u < 1. \end{cases}$$

where

$$Q_i^*(u) = Q_i[F_i(\text{domain lower limit}) + (F_i(\text{domain upper limit}) - F_i(\text{domain lower limit}))u]$$

The employment of composite models in dealing with real loss data can be found in a series of papers; Cooray and Ananda (2005), Scollnik and Sun (2012), Nadarajah and Bakar (2014); Bakar et. al. (2015) among others.

Folded model

The density of the folded model is given by

$$f(x) = g(x) + g(-x), \quad x > 0$$

where g(x) is the parent distribution whose support is of real values.
The inverse function is given by

$$F^{-1}(p) = G^{-1}\left(\frac{1+p}{2}\right)$$

where $G^{-1}(\cdot)$ is the cdf of the parent distribution.

- Brazauskas and Kleefeld (2011) has analyzed some folded and log folded models with application to the Norwegian fire loss data. A new scaled folded t model is found promising in this context.
- Later, Nadarajah and Bakar (2015) proposed several new folded models and derive their properties. At least one model has a better performance than the previously introduced.

Skewed symmetric model

The density of the skewed symmetric model is given by

$$f(x) = 2h(x)G(x) \quad x > 0$$

where $h(x)$ and $G(x)$ are the pdf and cdf of the parent distribution, not necessarily belongs to similar distribution. The model however does not have any closed form in general.

- In the context of insurance loss modelling, skewed-normal and skewed-student distribution have been analyzed on Danish and US indemnity losses by Eling (2012).
- The authors remarks that both models reasonably competitive in describing insurance data as well as tail risk measures.

Arc tan model

The density of the arc tan model is given by

$$f(x) = \frac{1}{\arctan(\alpha)} \frac{\alpha g(x)}{1 + (\alpha(1 - G(x))^2} \quad x > 0$$

where $g(x)$ and $G(x)$ are the pdf and cdf of the parent distribution, respectively. The arc tan model has closed form for its cdf and inverse cdf.

Risk Measures

Value-at-Risk

The empirical estimate of VaR can be obtained as:

$$VaR_X(\alpha) = \widehat{F}^{-1}(\alpha),$$

where $\widehat{F}(\cdot)$ denotes the empirical cdf.
The theoretical estimates of VaR can be obtained using the formula:

$$VaR_X(\alpha) = \begin{cases} F_1^{-1}\left(\alpha(1+\phi)F_1(\theta)\right), & \text{if} \quad 0 < \alpha \leq \frac{1}{1+\phi}, \\ F_2^{-1}\left(F_2(\theta) + (\alpha(1+\phi)-1)\left(1-F_2(\theta)\right)/\phi\right), & \text{if} \quad \frac{1}{1+\phi} < \alpha < 1. \end{cases}$$

Conditional Tail Expectation

The empirical estimate of CTE can be obtained as:

$$CTE_X(\alpha) = \frac{1}{1-\alpha}\int_{\alpha}^{1} \widehat{F}^{-1}(s)ds,$$

where $\widehat{F}(\cdot)$ denotes the empirical cdf.
The theoretical estimates of CTE can be obtained using the formula:

$$CTE_X(\alpha) = \frac{1}{1-\alpha}\int_{\alpha}^{1} VaR_X(s)ds.$$

Computer Implementation with R

- The R implementation of the generated models for computing the pdf, cdf, qf and random generated values can be handled by the `do.call` function. The proposed method accepts combination of any two arbitrary distributions to form the parent distributions.
- The `do.call` function allows calling any function in R with its argument specified using a list.
- For instance, the Weibull function can be called by `do.call( "weibull", list=(shape=2,scale=1) )`
- An R package `gendist` has been develop to summarize all these functions.

Estimation

- Below is a practical example for finding the estimated value of parameters of the composite Weibull-Paralogistic model using the maximum likelihood method with the `gendist` package.
- The density of the composite model is called by the function `dcomposite(...)` with Weibull and Paralogistic parameters being its arguments.

1. First, create the negative log-likelihood function for a general composite model:

```
R> nloglik <- function(p, spec1, arg1, spec2, arg2){
+ tt <- 1.0e20
+ if(all(p>0)){
+ tt <- -sum(log(dcomposite(x, spec1, arg1, spec2, arg2)))
+ }
+ return(tt)
+ }
```

2. Next, minimize the function in order to obtain the optimal solution:

```
R> par <- nlm(function(p){nloglik(p,spec1 = "weibull",
+ arg1 = list(scale = p [1]),
+ spec2 = "paralogis", arg2 = list(scale = p [2]))}, p = c(1,1))$estimate
```

Simulation

- In what follows, we describe the simulation of the arc tan model with parent distribution of Weibull (shape parameter: $\beta = 2$ and scale parameter: $\lambda = 0.5$)

1. Ten thousand sample sizes are generated from the inverse function of the arc tan model.

```
R> nsim <- 10000
R> nsiz <- 100
R> est1 <- matrix(0,nsiz,nsim)
R> mm1 <- matrix(0,nsiz,nsim)
R> bias <- rep(0,nsiz)
R> mse <- rep(0,nsiz)
R> ss <- rep(0,nsiz)
R> alpha <- 1.5
```

2. Estimate the parameter $\hat{\alpha}_i$ for $i = 1, 2, ..., 10000$.
3. Bias and mean squared error are obtained using

$$bias_{\alpha}(n) = \frac{1}{10000} \sum_{i=1}^{10000} (\hat{\alpha}_i - \alpha)$$

and

$$MSE_{\alpha}(n) = \frac{1}{10000} \sum_{i=1}^{10000} (\hat{\alpha}_i - \alpha)^2.$$

```
R> for (j in 1:nsiz){
for (i in 1:nsim){
x <-rarctan(j*10, alpha=alpha, spec="weibull", arg=c(shape=2,scale=0.5))\\
nlogl <-function(p, alpha, spec, arg)
{
tt <- 1.0e20
if(all(p>0)){
tt <- -sum(log( darctan(x, alpha, spec, arg) ))\\
}
return(tt)
}
est <-nlm(function(p) nlogl(p, alpha=p, spec="weibull",
arg=list(shape=2,scale=0.5)), p=1, hessian=T )
est1[j,i] <-est\$estimate[1]
mm1[j,1] <-solve(est\$hessian)[1,1]
}
bias[j] <-  mean(est1[j,] - alpha)
mse[j] <-mean((est1[j,] - alpha)^ 2)
ss[j] <-mean(mm1[j,]\^\ (1/2))
}
```

References

- Cooray, K., & Ananda, M. M. (2005). Modeling actuarial data with a composite lognormal-Pareto model. Scandinavian Actuarial Journal, 2005(5), 321-334.
- Bakar, S. A., Hamzah, N. A., Maghsoudi, M.,& Nadarajah, S. (2015). Modeling loss data using composite models. Insurance: Mathematics and Economics, 61, 146-154.
- Bakar, S. A. A., Nadarajah, S., Adzhar, Z. A. A. K., & Mohamed, I. (2016). Gendist: An R Package for Generated Probability Distribution Models. PloS one, 11(6).
- Brazauskas, V., & Kleefeld, A. (2011). Folded and log-folded-t distributions as models for insurance loss data. Scandinavian Actuarial Journal, 2011(1), 59-74.
- Gomez-Deniz, E., & Caldern-Ojeda, E. (2015). Modelling insurance data with the Pareto ArcTan distribution. ASTIN Bulletin: The Journal of the IAA, 45(3), 639-660.
- Miljkovic, T., & Grn, B. (2016). Modeling loss data using mixtures of distributions. Insurance: Mathematics and Economics, 70, 387-396.