Communications for Statistical Applications and Methods
2016, Vol. 23, No. 1, 85–91

http://dx.doi.org/10.5351/CSAM.2016.23.1.085
Print ISSN 2287-7843 / Online ISSN 2383-4757

Efficient simulation using saddlepoint approximation for aggregate losses with large frequencies

Jae-Rin Cho[a], Hyung-Tae Ha[1,b]

[a]Korea Insurance Research Institute, Korea
[b]Department of Applied Statistics, Gachon University, Korea

Abstract

Aggregate claim amounts with a large claim frequency represent a major concern to automobile insurance companies. In this paper, we show that a new hybrid method to combine the analytical saddlepoint approximation and Monte Carlo simulation can be an efficient computational method. We provide numerical comparisons between the hybrid method and the usual Monte Carlo simulation.

Keywords: large claim frequency, aggregate claim amount, saddlepoint approximation, simulation

1. Introduction

Compounding distributions are a major concern in insurance risk models to model the aggregate claim amounts in a fixed period of time. The accurate and efficient computation for the distribution of the total aggregate loss should be a mandatory step to determine managerial issues (such as the optimal premium and reservoir) for insurance companies. Closed-form solutions of the distribution of the collective risk models are not available in most cases; however, recursive methods for certain classes such as Panjer class have been discussed to calculate exact quantities. Since the recursive methods are often extremely time consuming, computational methods such as Monte Carlo simulation have extensively been discussed, which become even more competitive due to the progression of modern computer processing power.

In this paper, we are interested in aggregate claim amounts with a large number of claims, which often occur in car insurance policies. For example, the expected number of claims is 10,000 for a car insurance company that has 0.1 million policy holders with 10% accident probability during a certain period of time. In such circumstances, even Monte Carlo simulation may not be very efficient to calculate the distributions or quantiles. Some researchers such as Embrechts *et al.* (1985), Gatto (2010), and Jensen (1991), utilized analytical approximation methods such as saddlepoint approximation to solve this inefficiency. Whereas the saddlepoint approximation is known to be computationally efficient in addition to providing very accurate tail probabilities, calculation for some distributional quantities such as quantiles still remain challenging because the inversion of the saddlepoint distribution approximation is failed in most cases.

We introduce a hybrid computational method to combine the benefits from two different methods of Monte Carlo simulation and saddlepoint approximations to achieve both quantile computation

[1] Corresponding author: Department of Applied Statistics, Gachon University, 1342 Sungnamdae-ro, Sujung-gu, Sungnam-ci, Kyunggi-do 13120, Korea. E-mail: htha@gachon.ac.kr

Published 31 January 2016 / journal homepage: http://csam.or.kr

and computational efficiency for aggregate insurance losses with large frequencies that were originally discussed in McLeish (2014). Daniels (1954) and Lugannani and Rice (1980) versions of the saddlepoint approximation will be utilized as a scheme to determine acceptance-rejection regions. Numerical comparisons in Section 3 admit that the hybrid method can be a viable solution to efficiently provide quantiles of compounding distributions with large frequencies.

2. Simulation using saddlepoint approximation

We first introduce a collective risk model. Let Z be the total claim amounts occurred in an insurance portfolio within a given period of time, which is a sum of a stochastically determined number $N(t)$ of independent random variables X_i with distribution function F_i. The collective risk model can be expressed as

$$Z = X_1 + \cdots + X_{N(t)},$$

where claim frequency $N(t)$ is a discrete positive random variable with $p_n = \Pr(N(t) = k)$, $k = 0, 1, \ldots$ and X_i is a continuous random variable with probability density function f_i. There is a finite probability of no loss occurring over the considered time period if $N(t) = 0$ is allowed, i.e. $\Pr(N(t) = 0) = \Pr(Z = 0)$. $N(t)$ and the single of random variables X_i are independent for all i.

The distribution of the aggregate loss, denoted as $H(\ell)$, can be expressed in terms of convolution formula as

$$\begin{aligned} H(\ell) = \Pr(Z \le \ell) &= \sum_{k=0}^{\infty} \mathbf{Pr}(Z \le \ell | N(t) = k) \mathbf{Pr}(N(t) = k) \\ &= \sum_{k=0}^{\infty} p_k F^{(k)*}(\ell), \end{aligned}$$

where $F^{(k)*}(\ell) = \mathbf{Pr}(X_1 + \cdots + X_k \le \ell)$ is the k^{th}-convolution formula of $F(\cdot)$ calculated recursively from $(k-1)^{th}$-convolution as

$$F^{(k)*}(\ell) = \int_0^{\ell} F^{(k-1)*}(\ell - x) f(x) dx$$

with

$$F^{(0)*}(\ell) = \begin{cases} 1, & \ell \ge 0, \\ 0, & \ell < 0. \end{cases}$$

Though the obtained formula is analytic, its direct calculation is involves many integrations, such that the computations in practice are extremely heavy. Whereas the convolution are available in closed-form only in special cases, its moment generating function, denoted by $\chi(s)$, can be simply expressed as

$$\chi(s) = \sum_{i=0}^{\infty} (\varphi(s)) p_i = \psi(\varphi(s)),$$

where the moment and probability generating functions of random variables X and $N(t)$ are respectively $\varphi(s) = E[e^{sX}]$, and $\psi(k) = E[k^{N(t)}] = \sum_{i=0}^{\infty} k^i p_i$.

Once the moment generating function of an aggregate claim distribution is well defined in closed functional form, a saddlepoint approximation may be employed to provide accurate approximate density and distribution functions. On letting the cumulant generating function $\mathcal{K}(s) = \log[\chi(s)]$, for $s \in \mathcal{D} = (-t_0,\ t_1)$, where the domain $\mathcal{D}$ is a possibly semi- or infinite interval with $-t_0 < 0 < t_1$. When a unique solution ($t = \hat{t}$) is obtained from equating $\mathcal{K}^{(1)}(s) = x$ over the support, where $\mathcal{K}^{(i)}(s)$ is the i^{th} derivative of the cumulant generating function, Daniels (1954) version of the saddlepoint density approximation (SPA) is

$$f_{SP}(x) = \left(2\pi\,\mathcal{K}^{(2)}\left(\hat{t}\right)\right)^{-\frac{1}{2}} \exp\left(\mathcal{K}\left(\hat{t}\right) - x\hat{t}\right).$$

Normalization of the saddlepoint density approximation is required because the saddlepoint density approximant does not sum to unity. The normalized saddlepoint density approximant is often more accurate than the unnormalized saddlepoint density approximant. On denoting the normalizing factor by $\eta = (\int_0^\infty f_{SP}(x)dx)^{-1}$, the normalized saddlepoint density approximant (N-SPA) can be obtained as

$$f_{NSP}(x) = \eta f_{SP}(x).$$

The Lugannani and Rice (1980) version of saddlepoint distribution approximation can directly approximate tail probabilities $\mathbf{Pr}(Z \geq v)$, that is,

$$\mathbf{Pr}(Z \geq v) \approx 1 - \Phi\left(\hat{w}\right) + \phi\left(\hat{w}\right)\left(\frac{1}{\hat{u}} - \frac{1}{\hat{w}}\right),$$

where $\hat{w} = \sqrt{2(\hat{s}v - \mathcal{K}(\hat{s}))}\ \mathtt{sgn}(\hat{s})$, $\hat{u} = \hat{s}\sqrt{\mathcal{K}^{(2)}(\hat{s})}$, $\mathtt{sgn}(\hat{s}) = \pm 1, 0$ if $\hat{s}$ is positive, negative, or zero, and $\phi(\cdot)$ is the standard normal density function and $\Phi(\cdot)$ is the corresponding cumulative distribution function. This Lugannani and Rice formula is undefined if $\hat{u} = \hat{w} = 0$, which occurs when $v = E(Z)$ or $\hat{s} = 0$. The approximation in such case should be

$$\mathbf{Pr}(Z \geq v) \approx \frac{1}{2} - (2\pi)^{-\frac{1}{2}}\left(\frac{1}{6}\mathcal{K}^{(3)}(0)\left(\mathcal{K}^{(2)}(0)\right)^{-\frac{3}{2}} - \frac{1}{2}\left(\mathcal{K}^{(2)}(0)\right)^{-\frac{1}{2}}\right).$$

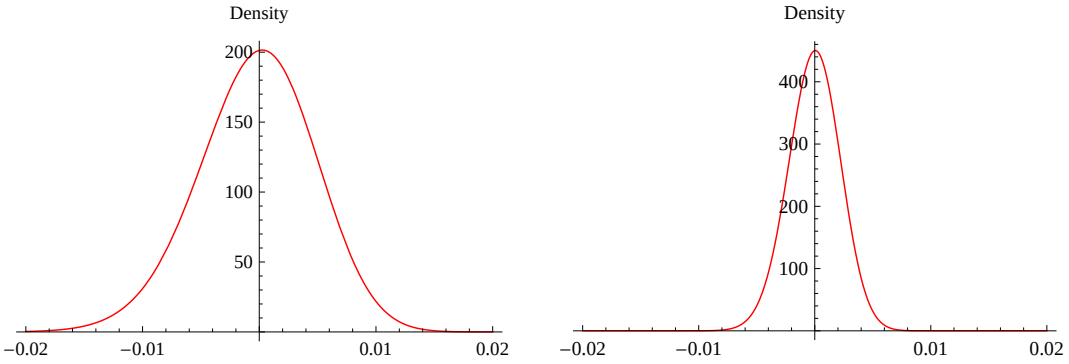
Now, we let $g(x)$ be the inverse function of $\mathcal{K}^{(1)}(s) = x$ so that the solution above is $\hat{s} = \mathcal{K}^{(1)^{-1}}(x) = g(x)$. Consider a transformed random variable $Y = g(X)$. From the transformation of variable techniques, the saddlepoint density and distribution approximants of Y are respectively

$$\begin{aligned} f_Y(y) &= f_{SP}(x)\left|\frac{\partial x}{\partial y}\right| \\ &= (2\pi)^{-\frac{1}{2}}\left(\mathcal{K}^{(2)}(y)\right)^{\frac{1}{2}} \exp\left(\mathcal{K}(y) - \mathcal{K}^{(1)}(y)y\right), \end{aligned} \tag{2.1}$$

and

$$\begin{aligned} \mathbf{Pr}(Y \leq v) &= \mathbf{Pr}\left(X \leq \mathcal{K}^{(1)}(v)\right) \\ &= \Phi(w(y)) - \phi(w(y))\left(\frac{1}{u(y)} - \frac{1}{w(y)}\right), \end{aligned} \tag{2.2}$$

where $w(y) = \sqrt{2(yv - \mathcal{K}(y))}\ \mathtt{sgn}(y)$ and $u(y) = y\sqrt{\mathcal{K}^{(2)}(y)}$. We use acceptance-rejection to generate random numbers under the saddlepoint density (2.3). We require a simple function to cover the saddlepoint density of Y, which can usually be a constant multiple of a typical continuous probability density function, from which random numbers can easily be generated. We call the covering simple function *dominating density*. Then, the area of saddlepoint density and the area between dominating density and saddlepoint density are used as acceptance and rejection regions, respectively.



(a) $\lambda_t = 100$, $\theta = 1/4$ and $\alpha = 80$ (b) $\lambda_t = 500$, $\theta = 1/4$ and $\alpha = 80$

Figure 1: *The saddlepoint density function for* $T = g(Y)$.

3. Aggregate losses with large frequencies: numerical examples

Suppose we want to simulate values of the random variables $Z = \sum_{i=0}^{N(t)} X_i$ where the frequency random variable $N(t)$ follows a Poisson distribution with parameter λ_t and X follows an independent gamma random variable with parameters θ and α. Then

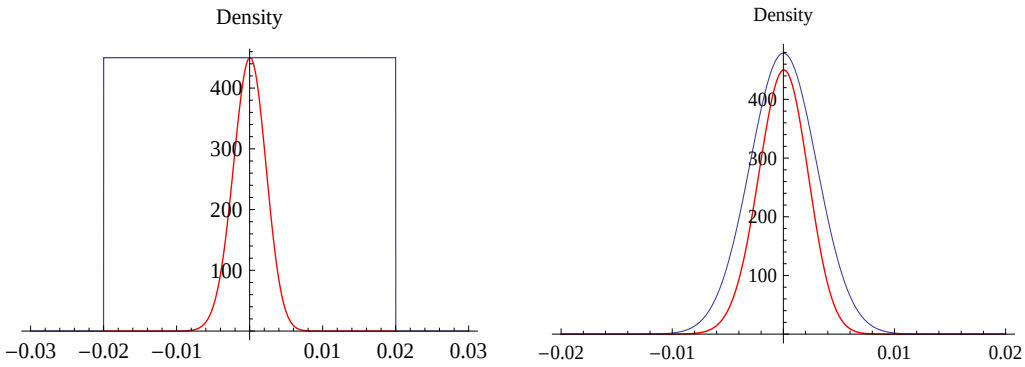
$$\chi(s) = \sum_{i=0}^{\infty} (\varphi(s))^i \frac{e^{-\lambda_t}\lambda_t^i}{i!} = \exp(\lambda_t \varphi(s) - \lambda_t),$$

where $\varphi(s) = (1-\theta s)^{-\alpha}$. Therefore, the cumulant generating function and its first and second derivatives of Z are respectively

$$\mathcal{K}(s) = \lambda_t(1-\theta s)^{-\alpha} - \lambda_t, \quad \mathcal{K}^{(1)}(s) = \alpha\theta\lambda_t(1-\theta s)^{-\alpha-1} \quad \text{and} \quad \mathcal{K}^{(2)}(s) = \alpha(\alpha+1)\theta^2\lambda_t(1-\theta s)^{-\alpha-2}.$$

Fortunately, the exact expression for the saddlepoint obtained by equating the equation $\mathcal{K}^{(1)}(s) = x$ can be determined as $\hat{s} = \theta^{-1}[1 - \{x/(\alpha\theta\lambda_t)\}^{-1/(\alpha+1)}]$, that is, the inverse function $g(x) = \theta^{-1}[1 - \{x/(\alpha\theta\lambda_t)\}^{-1/(\alpha+1)}]$. Therefore, we consider a new random variable $\mathcal{K}^{(1)}(Y) = X$, that is, $Y = \theta^{-1}[1 - \{X/(\alpha\theta\lambda_t)^{-1/(\alpha+1)}]$. The important step for this efficient simulation using saddlepoint approximation is to determine the dominating density for acceptance-rejection. We utilize constant multiples of two simple distributions, uniform and normal distributions, to cover most of the range of the probability density function of Y. It should be noted that the two simple cannot cover the entire range of the saddlepoint density function of Y since the saddlepoint density of Y has heavy tails. But, in practice, if probability of saddlepoint density corresponding to the covered range is close to unity, that is, dominating density covers a wide range of the saddlepoint density, the random number generation may be minimally influenced from the uncovered tail.

Figure 1 shows the saddlepoint densities; the left panel for the case of $\lambda_t = 100$, $\theta = 1/4$ and $\alpha = 80$, and the right panel for the case of $\lambda_t = 500$, $\theta = 1/4$ and $\alpha = 80$. Obviously, its variance becomes smaller when λ_t increases. Figure 2 shows the dominating function for $Y = g(X)$ with $\lambda_t = 500$, $\theta = 1/4$ and $\alpha = 80$; the left panel for the case of uniform distribution with a support $(-0.02,\ 0.02)$ and the left panel for the case of normal distribution with mean 0 and standard deviation 0.03. The multiplicative constants are 450.005 and 3.6 for uniform and normal dominating densities, respectively.



(a) 450.005 × Uniform distribution with a support (−0.02, 0.02)

(b) 3.6 × Normal distribution with mean 0 and standard deviation 0.03

Figure 2: *The dominating function for* $T = g(Y)$ *with* $\lambda_t = 500$, $\theta = 1/4$ *and* $\alpha = 80$.

Table 1: Computation time in seconds

	Case	RN			
		1,000	10,000	100,000	1,000,000
Computation time with uniform dominating function	$\lambda_t = 100$	2.652	4.089	17.596	153.396
	$\lambda_t = 500$	2.682	4.882	25.459	231.208
	$\lambda_t = 1{,}000$	2.777	4.682	25.068	232.347
	$\lambda_t = 10{,}000$	2.791	4.805	25.692	241.958
Computation time with normal dominating function	$\lambda_t = 100$	2.761	3.759	18.751	121.946
	$\lambda_t = 500$	2.605	3.589	14.822	110.276
	$\lambda_t = 1{,}000$	2.543	3.510	13.042	109.139
	$\lambda_t = 10{,}000$	2.497	3.555	13.073	110.386
Computation time using Monte Carlo simulation	$\lambda_t = 100$	6.150	39.050	313.080	3418.530
	$\lambda_t = 500$	19.640	154.062	1743.420	14692.600
	$\lambda_t = 1{,}000$	36.456	334.592	2853.410	29493.000
	$\lambda_t = 10{,}000$	313.594	3341.010	28844.800	284931.000

RN = Random Numbers.

Now, we compare the efficiency and accuracy of the proposed method and the usual Monte Carlo simulation. Consider various cases of large claims $\lambda_t = 100, 500, 1{,}000, 10{,}000$ when $\theta = 1/4$ and $\alpha = 80$. Table 1 compares the computational times between the proposed methods with uniform and normal dominating functions, and Monte Carlo simulation. Both proposed methods using uniform and normal dominating functions are extremely fast with comparison to the usual Monte Carlo simulation (Table 1). For instance, in order to generate 0.1 million numbers when $\lambda_t = 10{,}000$, the proposed method with uniform and normal dominating functions spends 232.347 and 109.139 seconds, whereas the Monte Carlo Simulation requires 29493. However, the proposed methods, using uniform and normal dominating functions, are more efficient than Monte Carlo Simulation by 127 and 270 times, respectively. The random number generation from uniform distribution is faster than normal distribution; however, the proposed method using normal dominating functions is more efficient than one using uniform dominating functions because the normal dominating function significantly reduces the rejection region. Table 2 shows the quantiles for the various cases. It is seen that those three methods provide similar estimates for the 90%, 95% and 99% quantiles. RN represents the number of generated random numbers. The numerical comparison was conducted on a personal computer

Table 2: Accuracy with uniform and normal functions and simulation

RN	Cases	λ_t			
		100	500	1,000	10,000
1,000	90%	2257.43	10474.9	20627.5	199312
		2285.81	10547.5	20773.2	202485
		2251.04	10542.1	20844.6	202534
	95%	2318.46	10660.7	20876.1	201042
		2353.89	10770.0	21032.9	203249
		2312.60	10750.6	21079.4	203304
	99%	2483.62	11059.9	21240.3	203334
		2506.41	11107.1	21454.8	204462
		2495.39	11028.3	21530.7	205243
10,000	90%	2258.01	10489.8	20543.8	199447
		2261.31	10573.2	20825.7	202659
		2262.98	10575.9	202450	202560
	95%	2331.34	10666.3	20823.5	201001
		2337.07	10743.2	21064.4	203420
		2349.65	10731.4	203107	203275
	99%	2479.65	10993.3	21331.2	202952
		2474.01	11047.4	21514.3	204615
		2493.56	11048.8	204376	204759
100,000	90%	2259.20	10485.0	20530.1	199282
		2258.09	10577.6	20814.8	202572
		2259.02	10579.6	20823.1	202567
	95%	2336.90	10672.0	20815.7	200934
		2334.76	10744.5	21046.5	203312
		2335.81	10747.1	21055.1	203313
	99%	2485.30	10998.0	21319.7	203066
		2476.82	11047.6	21491.5	204678
		2480.68	11060.4	21483.3	203938
1,000,000	90%	2260.56	10478.6	20530.3	199274
		2260.07	10578.6	20818.4	202579
		2260.54	10759.4	20817.9	202577
	95%	2337.20	10662.6	20814.7	200937
		2336.59	10746.7	21052.6	203306
		2336.95	10746.5	21052.1	203307
	99%	2482.09	10998.5	21315.8	203060
		2482.51	11061.3	21494.0	204687
		2482.65	11062.6	21496.4	203940

RN = Random Numbers.

with a Core TM i5-4210M CPU 2.6 GHZ.

4. Concluding remarks

Saddlepoint approximation was utilized to construct acceptance-rejection algorithm to generate random numbers for compounding sums. The choice of dominating density shall be carefully decided because the performance of the proposed simulation depends on the simplicity and closeness of the dominating density to the saddlepoint approximation of the aggregate claim distribution. In addition, simple dominating density for acceptance-rejection such as uniform or normal distributions can often provide accurate quantiles of compounding sums that are extremely efficient in computation. The proposed hybrid methods can be considered as a viable alternative to obtain the probabilistic quantities of compounding sums with large claim frequencies. It will be interesting to study the analytical and computational properties of an appropriate dominating density under various future circumstances.

Acknowledgements

This research was supported by the Gachon University research fund of 2014 (GCU-2014-0172). The authors wish to express sincere thanks to the editor and two anonymous referees.

References

Daniels HE (1954). Saddlepoint approximations in statistics, *Annals of Mathematical Statistics*, **25**, 631–650.

Embrechts P, Jensen JL, Maejima M, and Teugels JL (1985). Approximations for compound Poisson and Plya processes, *Advances in Applied Probability*, **17**, 623–637.

Gatto R (2010). A saddlepoint approximation to the distribution of inhomogeneous discounted compound Poisson processes, *Methodology and Computing in Applied Probability,* **12**, 533–551.

Jensen JL (1991). Saddlepoint approximations to the distribution of the total claim amount in some recent risk models, *Scandinavian Actuarial Journal*, **1991**, 154–168.

Lugannani R and Rice SO (1980). Saddlepoint approximation for the distribution of the sum of independent random variables, *Advances in Applied Probability*, **12**, 475–490.

McLeish D (2014). Simulating random variables using moment-generating functions and the saddlepoint approximation, *Journal of Statistical Computation and Simulation*, **84**, 324–334.

Received July 15, 2015; Revised January 5, 2016; Accepted January 5, 2016