Importance Sampling for Sums of Lognormal Distributions, with Applications to Operational Risk

Marco Bee

Importance Sampling for Sums of Lognormal Distributions, with Applications to Operational Risk

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2008-0113
Manuscript Type:	Original Paper
Date Submitted by the Author:	14-May-2008
Complete List of Authors:	Bee, Marco; University of Trento, Department of Economics
Keywords:	Defensive Mixtures, Importance Sampling, Cross-Entropy, Tail Probability, Compound Distributions
Abstract:	In this paper we use Importance Sampling to estimate tail probabilities for a finite sum of lognormal distributions. We use a defensive mixture, and develop a method of choosing the parameters via the EM algorithm; we also consider the technique which assumes the importance sampling density to belong to the same parametric family of the random variables to be summed. In both cases, the instrumental density is found by minimizing Cross-Entropy. A comparison based on several simulation experiments shows that the defensive mixture has the best performance. Finally, we study the Poisson-lognormal compound distribution framework and present a real-data application
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
Style_latex.tex bee.zip	

Importance Sampling for Sums of Lognormal Distributions, with Applications to Operational Risk

Marco Bee

*Department of Economics, University of Trento**

Running head: Importance Sampling for Sums of Lognormals.

*Postal address: via Inama 5, 38100 Trento - Italy. Phone: +39-0461-882296. Fax: +39-0461-882222. E-mail: marco.bee@economia.unitn.it

1 Introduction

In the last three decades, Monte Carlo (MC) simulation has become a very popular tool. This success can be traced back to at least two reasons: first, the widespread availability of cheap computing power resulted in more feasible execution times; second, the development of models where no analytical solutions exist has forced researchers to resort to computer-intensive methodologies. The latter remark also applies to the case when only asymptotic approximations exist, and the small sample behavior of estimators or test statistics must be investigated by means of stochastic simulation.

In some cases, however, standard MC is not the most appropriate tool; this happens, for example, when we deal with rare events. Suppose that one were interested in the estimation of the probability $p = P(X \geq c)$ for some "large" threshold c. The Crude MC (CMC) approach consists in simulating N observations $x_1, \ldots, x_N$ from the distribution of X and computing the estimate

$$\hat{p}^{MC} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \geq c\}}, \tag{1}$$

where $\mathbf{1}_A$ is the indicator function of the set A. If c is the α quantile of the distribution, we expect $\sum_{i=1}^{N} \mathbf{1}_{\{X_i \geq c\}}$ to contain only $N \cdot (1-\alpha)$ non-zero summands, so that, for very large values of α, N must be very large to obtain an estimator with good properties. Formally, this can be seen by means of the standard efficiency measure for estimators of rare event probabilities, namely the relative error $\tau =: \sqrt{\mathrm{var}(\mathbf{1}_{\{X_i \geq c\}})}/E(\mathbf{1}_{\{X_i \geq c\}})$; see, for, example, Asmussen and Binswanger (1997). For the CMC estimator the relative error diverges as $p \to 0$:

$$\tau = \frac{\sqrt{p(1-p)}}{p} \approx \frac{1}{\sqrt{p}} \to \infty \text{ as } p \to 0.$$

The difficulty becomes even more relevant when we are interested in estimating the moments of the conditional distribution of X, as the following example shows.

Example 1. Consider the estimation of $\mu_c = E(X|X \geq c)$ when X has the standard lognormal distribution, i.e. $X \sim \mathrm{Logn}(0,1)$. The CMC estimator is given by

$$\hat{\mu}_c^{MC} = \frac{1}{\#\{X_i \geq c\}} \sum_{i=1}^{N} X_i \mathbf{1}_{\{X_i \geq c\}}, \tag{2}$$

where $X_1, \ldots, X_N$ is a random sample from X. Figure 1 shows the MSE of $\hat{\mu}_c^{MC}$ as a function of c, obtained by repeating $B = 10000$ times the following two steps:

1. Simulate $N = 10^6$ standard lognormal random numbers; notice that such a large value of N is necessary in order to get a reasonable estimate of the variance for the largest values of c considered in the example;

2. Use (2) to compute the CMC estimate of μ_c.

Then, we compute the mean and the variance, respectively given by $\hat{\bar{\mu}}_c = \sum_{i=1}^{B} \hat{\mu}_c^{(i)}/B$ and $\text{var}(\hat{\mu}_c) = \sum_{i=1}^{B}(\hat{\mu}_c^{(i)} - \hat{\bar{\mu}}_c)^2/B$, where $\hat{\mu}_c^{(i)}$ is the i-th estimate ($i = 1, \ldots, B$) obtained in the simulation procedure; finally, the bias is equal to $b(\hat{\mu}_c) = \hat{\bar{\mu}}_c - \mu_c$. In the lognormal case we know from elementary probability theory the true conditional expectation:

$$\mu_c = E(X|X \geq c) = e^{\mu+\sigma^2/2} \cdot \frac{1 - \Phi(\xi - \sigma)}{1 - \Phi(\xi)},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the density and the distribution functions of the standard normal distribution and $\xi = (\log(c) - \mu)/\sigma$. It follows that in this example $E(X|X \geq c) = \exp\{1/2\}(1 - \Phi(\log(c) - 1))/(1 - \Phi(\log(c)))$.

FIGURE 1 HERE

It can be seen that the MSE increases sharply as c gets larger. As for the variance of the estimator, it is very close to the MSE, thus no graphical representation is given. The latter result is not surprising in view of the fact that $\text{bias}^2 = O(1/N)$, whereas $\text{variance} = O(1/N^2)$, so that in simulations bias^2 is small if N is large enough. ∎

Importance Sampling (Hammersley and Handscomb 1964; see Casella and Robert 2004, sect. 3.3 or Glasserman 2003, sect. 4.6 for reviews) is a very powerful variance reduction technique. The setup is based on standard MC, and Importance Sampling (IS) shares many desirable properties with it; the main difference is that IS does not simulate observations from the probability measure of interest P, but from an instrumental probability measure $\tilde{P}$ which assigns "more weight" to the event of interest. Not surprisingly, the features of $\tilde{P}$ are of crucial importance, because they can result in either large efficiency gains or poor quality estimators. Thus, IS is commonly considered one of the most effective but also most complex variance reduction techniques.

The theory of IS is well established for light-tailed distributions: the IS density is usually based on exponential tilting (more concretely, on the moment generating function), and efficiency results are known, at least asymptotically. On the other hand, when working with heavy-tailed distributions, no all-purpose recipe for finding the IS probability measure is available; moreover, from the theoretical point of view, the lack of exponential moments makes it difficult to develop limit results. Thus, the properties of the estimators must often be studied numerically.

In this article we derive an IS technique for the computation of tail probabilities when the distribution of interest is a sum of *iid* lognormal distributions. More precisely, the IS density used in this paper belongs to the Defensive Mixture class (Hesterberg 1985) and the optimal density is found by means of the minimum Cross-Entropy method. The main contribution of the paper is twofold. First, we develop a methodology which combines these two tools, namely Defensive Mixtures and Cross-Entropy minimization; second, a comparison with more standard IS techniques shows that the approach proposed here has very favorable properties. Technically, this procedure has marked similarities with maximum likelihood

estimation of the parameters of a mixture distribution, so that it can be tackled by means of the EM algorithm.

We also extend the IS solution found in this setup to the estimation of the quantiles of a compound Poisson-lognormal distribution. This model is frequently employed in the analysis of the total loss distribution in actuarial and operational risk applications. In the latter field, IS can be of paramount importance, because the computation of risk measures (also for regulatory prescriptions contained in the New Basel Accord on Banking Supervision; see Basel Committee on Banking Supervision 2005) requires the estimation of extreme quantiles, for which CMC suffers from the drawbacks shown in example 1 above.

The rest of the paper is organized as follows. Section 2 reviews the basic IS methodology and gives some details about heavy-tailed distributions; Section 3 uses the minimum Cross-Entropy approach to derive the parameters of the IS density for the estimation of the tail probability of a finite sum of lognormal distributions; Section 4 presents the results of several simulation experiments aimed at verifying the properties of the estimators; Section 5 applies the technique to a compound Poisson-lognormal distribution and computes tail probabilities in an operational risk setup; Section 6 concludes and discusses some directions for future research.

2 Importance Sampling for sums of lognormals

2.1 Preliminaries: basic Importance Sampling

Most of the time, MC simulation is devoted to the computation of a definite integral. Let X be a random variable defined on some probability space $(\Omega, \mathcal{F}, P)$ and assume that it is absolutely continuous with density $f(\cdot)$; moreover, let $h(\cdot)$ be a known function. Consider evaluating the following integral:

$$\mu = E_f(h(X)) = \int_{\mathcal{X}} h(x) f(x) dx. \tag{3}$$

Sometimes, the density of X is too complex, and the analytical evaluation of (3) is difficult or even impossible. A readily available solution is MC simulation, which consists in simulating $x_1, \ldots, x_N$ independently from $f(\cdot)$ and computing $\hat{\mu}^{MC} = (1/N) \sum_{i=1}^{N} h(x_i)$. The convergence of $\hat{\mu}^{MC}$ follows from the strong law of large numbers. The rate of decrease of the variance of the estimator is equal to $O(1/N)$, so that the variance is equal to σ^2/N for some σ. If σ is very large, the achievement of the desired precision level requires an extremely large N.

To introduce IS, note that (3) has the following alternative representation:

$$\mu = E_f(h(X)) = \int_{\mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x) dx,$$

where $g(\cdot)$ is also a density. It follows that

$$\mu = E_g \left(\frac{h(X) f(X)}{g(X)} \right), \tag{4}$$

where the expectation is taken with respect to the density $g(\cdot)$. Equation (4) provides another method of simulating X:

Algorithm 1 *(Importance Sampling)*

- *Simulate* $x_1, \ldots, x_N$ *independently from* $g(\cdot)$*;*
- *Compute*

$$\hat{\mu}_{IS} = \frac{1}{N} \sum_{i=1}^{N} h(x_i) \frac{f(x_i)}{g(x_i)}. \tag{5}$$

Intuitively, the instrumental distribution g "assigns a larger probability" (in a sense to be made more precise later) to the event of interest. The estimator $\hat{\mu}_{IS}$ is called an *importance sampling estimator* of μ, and $g(\cdot)$ is the *importance sampling density*. The ratio $r(x) = f(x)/g(x)$, usually referred to as the *likelihood ratio*, can be interpreted as a weight, so that (5) is a weighted mean.

Unbiasedness, consistency and asymptotic normality of the estimator $\hat{\mu}_{IS}$ follow from the asymptotic theory of standard MC, under the conditions that the support of g includes the support of f and that g is such that the variance of the estimator is finite. This has the interesting implication that the choice of g can be based on efficiency criteria; in particular, we could seek a density g^* which is easy to simulate and is such that $\text{var}_{g^*}(\hat{\mu}_{IS})$ is small. A necessary condition for the variance of (5) to exist is that

$$E_g \left(\frac{h^2(X) f^2(X)}{g^2(X)} \right) = E_f \left(\frac{h^2(X) f(X)}{g(X)} \right) = \int_{\mathbb{R}} h^2(x) \frac{f^2(x)}{g(x)} dx < \infty.$$

As pointed out by Casella and Robert (2004, sect. 3.3.2), this implies that we should analyze carefully the behavior of $r(X)$ in the tails of the distribution, because importance sampling distributions with an unbounded likelihood ratio are likely to give estimators with infinite variance and/or widely varying weights. Loosely speaking, "good" IS distributions g should have a thicker right tail than f. Two types of sufficient conditions for the finiteness of $E_f(h^2(X)f(X)/g(X))$ have been proved by Geweke (1989):

$$\frac{f(x)}{g(x)} < M \ \forall x \in \mathcal{X} \text{ and } \text{var}_f(h) < \infty; \tag{6}$$

$$\mathcal{X} \text{ is compact, } f(x) < F \text{ and } g(x) > \epsilon \ \forall x \in \mathcal{X}, \tag{7}$$

where F is a positive constant and ϵ is a small positive constant. As for the conditions (6), if the variance of the simple Monte Carlo estimate is finite without importance sampling, and if the likelihood ratio is bounded, then the variance with importance sampling is bounded. We use defensive mixture sampling (described below) to bound the likelihood ratio. However, it is worth noting that there are IS algorithms which satisfy neither (6) nor (7).

The function g that minimizes the variance of the estimator can be found explicitly (Casella and Robert 2004, Rubinstein and Kroese 2008). However, the result is of little help in applications as it requires the knowledge of $\int h(x)f(x)dx$, i.e. the integral we are interested in. Thus, the obvious question is: how do we choose the IS density? Ideally, we would like to find a standard procedure which works in all setups.

Solutions of this type, mainly based on tilted densities (Ross 2006), are often available when working with light-tailed distributions. For heavy-tailed distributions the problem has to be solved differently, mostly on a case-by-case basis.

2.2 Importance Sampling for sums of lognormal random variables

Given k *iid* random variables $X_1, \ldots, X_k$, the rate at which $p_k \stackrel{\text{def}}{=} P(X_1 + \cdots + X_k > c_k) \to 0$ for $c > E(X)$ and $c_k = kc$ has been thoroughly investigated by the theory of *Large Deviations* (Durrett 1996, sect. 1.9). Essentially all the results assume the existence of the moment generating function of X; see Mikosch and Nagaev 1998 for a review of the Large Deviations methodology in the heavy-tailed setup. Furthermore, the problem studied by Large Deviations is somewhat different from what is being investigated in this paper, because here we are concerned with fixed values of p_k and k, not with an asymptotic (as $p_k \to 0$) tail probability.

It is well known that, although the lognormal distribution has all finite moments, its moment generating function does not exist (see, for example, Moran 1984). As a consequence, both the IS procedure based on tilted densities and the theory of Large Deviations are of little help in this case; thus, different tools are needed both to find an IS density and to investigate the speed of convergence.

In this paper we tackle the problem by means of Defensive Mixtures (DM; Hesterberg 1995, Davison and Hinkley 1997, p. 457). When considering a single r.v., this approach builds the IS density as a mixture of the distribution of interest itself and of another distribution, usually, but not necessarily, belonging to the same parametric class. For example, in the setup of example 1, with $X \sim \text{Logn}(\mu, \sigma^2)$, a reasonable choice for the IS density is a mixture of X and another lognormal distribution with a larger expected value: $X_2 \sim \text{Logn}(\mu_2, \sigma_2^2)$, with $\mu_2 = \mu + t$, $t \in \mathbb{R}^+$. The IS density would therefore be given by:

$$g(x) = \pi f(x; \mu, \sigma^2) + (1 - \pi) f_2(x; \mu_2, \sigma_2^2), \qquad 0 < \pi < 1, \tag{8}$$

where f and f_2 are respectively the density of X and X_2.

In a multivariate setup, namely when we estimate the tail probability of $Y = \sum_{i=1}^k X_i$, a defensive mixture density is a mixture of the joint density:

$$g_k(\boldsymbol{x}) = \pi f_k(\boldsymbol{x}) + (1 - \pi) g_k(\boldsymbol{x}),$$

where g_k is some other k-dimensional density. Suppose now that f_k has independent marginals:

$$f_k(\boldsymbol{x}) = \prod_{i=1}^k f(x_i).$$

Then typically one would let g_k also have independent marginals:

$$g_k(\boldsymbol{x}) = \prod_{i=1}^k g(x_i).$$

If the X_i's are lognormal, the individual components $g(x_i)$ could be lognormal, or a mixture of lognormals, so that in general one would have the following compound mixture distribution:

$$g_k(\boldsymbol{x}) = \pi \prod_{i=1}^{k} f(x_i) + (1-\pi) \prod_{i=1}^{k} (\lambda g(x_i) + (1-\lambda) g_2(x_i)).$$

In this paper we restrict ourselves to the case $\lambda = 0$, so that with probability π we sample from a $\text{Logn}(\mu, \sigma^2)$ population and with probability $(1-\pi)$ from a $\text{Logn}(\mu_2, \sigma_2^2)$ population.

As pointed out by Hesterberg (1995), this approach has the advantage of providing weights bounded above by $1/\pi$; furthermore, it is quite clear that conditions (6) and (7) are satisfied. On the other hand, the main difficulty is that, in principle, three parameters (π, t and σ_2) have to be chosen.

2.3 An asymptotically efficient approach

The lognormal distribution is subexponential (Embrechts *et al.* 1997, sect. 1.3.2 or Asmussen *et al.* 2000, sect. 2.1) and heavy-tailed; more precisely, it belongs to the Maximum Domain of Attraction (MDA) of the Gumbel distribution (Embrechts *et al.* 1997, sect. 3.3). For the sum of random variables belonging to the intersection of the subexponential class and of the MDA of the Gumbel distribution, Asmussen *et al.* (2000) propose an IS density which satisfies the theoretical requirement of asymptotic efficiency (see e.g. Sadowsky 1993). The marginal density is of the form (Asmussen *et al.* 2000, p. 310)

$$g_A(x) = \begin{cases} \frac{\eta}{x(\log(x))^2} & x \in (a, +\infty), \\ \gamma l(x) & x \in [0, a], \end{cases} \tag{9}$$

where $a > e$ (where e is the base of natural logarithm), $\eta > 0$, l is an arbitrary strictly positive density on $[0, a]$ and γ is a normalizing constant; the joint density is the product of independent marginals. This distribution is very heavy-tailed, to the extent that for most choices of the parameters there is a relatively high probability of obtaining extremely large values: for example, with $\eta = 0.25$ and $a = 2.8$, the probability of observing a value larger than 10^6 is more than 1%, and for larger values of η this probability is even higher. Simulation experiments with 10000 replications demonstrated at least two overflows.

However, if the goal is estimating p_k, then one does not need to have g_A be so heavy-tailed beyond the threshold, as it would indeed be nearly optimal to have $g_A(x)/f(x)$ approximately constant for $y > c_k$. In fact, as noted by Asmussen and Kroese (2006, p. 550), the properties of the IS estimator obtained using (9) are rather poor for most parameter configurations. Intuitively, the variance of $\hat{p}^{IS}$ based on g_A is large because the distribution of the likelihood ratio is strongly concentrated around zero. As a result, the variance of the likelihood ratio is small, but the IS estimator $\hat{p}^{IS}$ is almost entirely determined by the few weights which are away from zero, so that it is highly unstable.

More formally, fix $\epsilon > 0$, let $h = \mathbf{1}_{\{x \geq c\}}$ and $A_\epsilon =: \{x_i : r(x_i) > \epsilon\}$. Then, if the IS density is such that one gets some fairly large weights in the region of interest, $\hat{p} \approx (1/N) \sum_{x_i \in A_\epsilon} h(x_i) r(x_i)$ with $x_i \sim g_A$. The relative error is given by $\tau = \sqrt{p_{A_\epsilon}(1 - p_{A_\epsilon})}/p$, whose magnitude is ultimately determined by p_{A_ϵ}: if p_{A_ϵ} is smaller than p, the performance can be even worse than CMC. The reason is that the

IS density has a much heavier tail than the target density, so that the simulation produces many large observations, for which both the likelihood ratio and the product $r(x)h(x)$ are essentially zero. Thus, the density (9) "puts more weight" than f on $[c, +\infty)$, but a non-negligible percentage of this weight is on the interval $[c_1, +\infty)$, where c_1 is such that $P(X > c_1) \approx 0$; this also explains why A_ϵ usually contains few observations.

In our opinion, in applications, an asymptotically bounded relative error is not the most relevant requirement of an IS estimator, because the aim consists mainly in the estimation of tail probabilities for small but fixed values of p, so that the limiting behavior of the estimator as $p \to 0$ is less important than the properties of the estimator for the true value of p, namely the value of p to be estimated. Therefore, we abandon the criterion of asymptotic efficiency and follow a different road.

2.4 Determining the optimal IS density

The most obvious approach to the selection of the parameters in (8) would be to choose numerical values of these parameters which minimize the variance of the IS estimator. As before, we assume to be interested in the estimation of p_k for some large threshold $c_k > E(Y)$; putting $h_k(\boldsymbol{x}) = \mathbf{1}_{\{\sum_{i=1}^k x_i > c_k\}}$, this probability can be written as

$$p_k = \int_{[0,\infty)^k} h_k(\boldsymbol{x}) f_k(\boldsymbol{x}) d\boldsymbol{x} = E_{f_k}(h_k(\boldsymbol{X})). \tag{10}$$

We must find a k-variate density g_k such that we simulate N iid vectors (of length k) $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_N^*$ from g_k and take $\hat{p}_k^{IS} = (1/N) \sum_{i=1}^N h_k(\boldsymbol{x}_i^*) r_k(\boldsymbol{x}_i^*)$ as an estimate of p_k, where the weight $r_k(\boldsymbol{x}) = f_k(\boldsymbol{x})/g_k(\boldsymbol{x})$ is the likelihood ratio.

Now minimizing the variance of $\hat{p}_k^{IS}$ is equivalent to minimizing the expected value $E_{f_k}(r_k(\boldsymbol{X})h_k(\boldsymbol{X}))$. To see why, notice that $\hat{p}_k^{IS}$ is the CMC estimator of $E_{g_k}(h_k(\boldsymbol{X})r_k(\boldsymbol{X}))$. We have:

$$\begin{aligned}
\text{var}(\hat{p}_k^{IS}) &= \frac{1}{N^2} \text{var}\left(\sum_{i=1}^N h_k(\boldsymbol{X}_i) r_k(\boldsymbol{X}_i) \right) = \frac{1}{N} \text{var}(h_k(\boldsymbol{X}) r_k(\boldsymbol{X})) = \\
&= \frac{1}{N} E_{g_k}[h_k(\boldsymbol{X})^2 r_k(\boldsymbol{X})^2] - p_k^2 \propto E_{g_k}(h_k(\boldsymbol{X})^2 r_k(\boldsymbol{X})^2) = \int_{[0,\infty)^k} h_k(\boldsymbol{X})^2 \frac{f_k(\boldsymbol{x})^2}{g_k(\boldsymbol{x})^2} g_k(\boldsymbol{x}) d\boldsymbol{x} = \\
&= \int_{[0,\infty)^k} h_k(\boldsymbol{X})^2 \frac{f_k(\boldsymbol{x})}{g_k(\boldsymbol{x})} f_k(\boldsymbol{x}) d\boldsymbol{x} = E_{f_k}(h_k(\boldsymbol{X})^2 r_k(\boldsymbol{X})) = E_{f_k}(h_k(\boldsymbol{X}) r_k(\boldsymbol{X})).
\end{aligned} \tag{11}$$

However, there are at least two difficulties with this approach. First, evaluating (11) can be complicated because it depends on both π and t. Second, and more important, the random variable $(1/N) \sum_{i=1}^N r_k(\boldsymbol{X}_i)$ with $\boldsymbol{X}_i \sim g$ may have a very large variance and/or be too concentrated around zero. Thus, minimizing (11) is not enough, because it implies minimization of the variance of the estimator but can produce an estimator whose convergence to the true value is too slow for practical purposes; namely, there are cases where consistency is just formal and the estimator is biased for any reasonable sample size. Therefore, an analysis of the whole distribution of $r_k(\boldsymbol{X})$ is necessary: for example, if we use the IS density (9), the variance of $r_k(\boldsymbol{X})$ is small but the variance of $\hat{p}$ is not.

Let's now focus on the density (8); for simplicity, we put $\sigma_2^2 = \sigma_1^2 = \sigma^2$, although the EM algorithm presented below (see sect. 3.1) could easily be extended to the estimation of σ_2^2 as well. The main reason for introducing such an IS density is that it guarantees bounded weights. This issue, which turns out to be crucial in the sum of lognormals case, is indeed negligible when working with a single lognormal distribution: taking $\pi = 0$ in (8), some straightforward algebra shows that

$$r(x) = \frac{\frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right\}}{\frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\log(x)-(\mu+t))^2}{2\sigma^2}\right\}} = \exp\left\{\frac{t^2 - 2\mu t - 2t\log(x)}{2\sigma^2}\right\}.$$

This is a monotonically decreasing function, and $\lim_{x\to 0} r(x) = +\infty$. Thus, $r(x)$ is unbounded when x is small, but the only x_i's which contribute to $\hat{p}^{IS}$ are those belonging to the set $A_c = \{x_i : x_i \geq c\}$, so that $r(x)$ is bounded for $x \in A_c$, and the upper bound is $r(c)$. Thus, in the single lognormal case, one can use an IS density of the form $\text{Logn}(\mu + t, \sigma^2)$ and minimize directly (11) with respect to t. However, in this setup one can integrate the density numerically, therefore we do not present further details here.

If $X_1, \ldots, X_k$ are independent, (10) can be rewritten as

$$p_k = E_{f_k}(h_k(\boldsymbol{X})) = \int_{[0,\infty)^k} h_k(x_1, \ldots, x_k) f(x_1) \cdots f(x_k) dx_1 \cdots dx_k.$$

If g_k has independent components, i.e. $g_k(\boldsymbol{x}) = \prod_{i=1}^k g(x_i)$, the IS approach is based on the fact that

$$p_k = E_{f_k}(h_k(\boldsymbol{X})) = E_{g_k}(h_k(\boldsymbol{X}) r_k(\boldsymbol{X})).$$

With these hypotheses, the likelihood ratio is equal to

$$r_k(\boldsymbol{x}) = \frac{\prod_{i=1}^k f(x_i)}{\prod_{i=1}^k g(x_i)}. \tag{12}$$

When $k > 1$, the distribution of $r_k(\boldsymbol{X})$ has a disturbing feature: the values of $r_k(\boldsymbol{x})$ which actually contribute to the estimator are those with $x_1, \ldots, x_k$ such that $\sum_{i=1}^k x_i > c$. It is clear that $\sum_{i=1}^k x_i$ can be larger than c even though one or more of the x_i's is arbitrarily small; but this implies that, if $\pi = 0$, $r_k(\boldsymbol{x})$ is unbounded. In addition to the aforementioned difficulties, the behavior of $r_k(\boldsymbol{X})$ as $k \to \infty$ is somewhat pathological, as summarized by the following theorem.

Theorem 1 *If both f_k and g_k have independent marginals, and under the condition*

$$E_{g_k}(|\log(f_k(\boldsymbol{X})/g_k(\boldsymbol{X}))|) < \infty,$$

the following results hold true:

$$E_{\boldsymbol{g}}\left(\frac{\prod_{i=1}^k f(X_i)}{\prod_{i=1}^k g(X_i)}\right) = 1 \text{ for any } k \in \boldsymbol{N};$$

$$\lim_{k\to\infty} \frac{\prod_{i=1}^k f(X_i)}{\prod_{i=1}^k g(X_i)} = 0 \text{ with } \tilde{P}\text{-probability 1.}$$

Proof. See Casella and Robert (2004, p. 551) or Glasserman (2003, p. 259).

The theorem has two implications. The first one, sometimes termed "weight degeneracy" (Casella and Robert 2004, p. 552) consists in the fact that, as k gets large, the distribution of the weights becomes more and more skewed, with most weights close to zero; in the limit an extremely large sample size is needed to get a single non-zero weight. However, the estimator is useless because it is essentially computed with only one observation. For intermediate values of k the estimator is downward biased unless N is very large. On the other hand, as π decreases, the maximum of the likelihood ratio gets larger: when using a defensive mixture, it is easy to see that $\max_{\boldsymbol{x}} r_k(\boldsymbol{x}) = \pi^{-1}$. The speed of convergence to zero is a decreasing function of π: as π gets small, there is an increasing probability of large values of the X_i's, for which the likelihood ratio is small.

Second, a naive optimization of (11) is unfeasible, because (i) two parameters have to be found (π and t), (ii) they also depend on k and c and (iii) numerical integration over a large dimensional space is problematic. Moreover, optimizing has its own costs; if the reduction of variance is approximately the same for all the numerical values of the parameter(s) in some interval(s), then it would be legitimate to choose any value(s) in the interval(s). Some numerical results on this issue will be given in section 4.

3 The Cross-Entropy approach

The crucial aspect of IS consists in finding the optimal instrumental density. This issue is actually twofold: one has to first choose the parametric form of the IS density, then to define an optimality criterion and use it for finding the parameters. In this section we propose the criterion of minimum Cross-Entropy (CE). In section 3.1 we apply it to the setup where the IS density belongs to the DM class, namely is a mixture of lognormals. Then we turn to the case where the IS density belongs to the same parametric family (i.e., lognormal) of the density of interest: in section 3.2 we implement the standard CE approach, in section 3.3 the Adaptive CE technique.

3.1 The Defensive Mixture approach

Hesterberg (1995) finds the numerical values of the parameters of the defensive mixture mostly on the basis of heuristic considerations. Here we develop a technique, based on the EM algorithm (Dempster *et al.* 1977), for determining the parameters according to the minimum CE approach. Thus, the parameters minimize the CE between the distribution of interest and the IS distribution, i.e. the defensive mixture.

The method of CE minimization was first proposed by Rubinstein (1997); see also Rubinstein and Kroese (2004) and Asmussen *et al.* (2005). Referring the interested reader to these references for details, we start by noting (Asmussen *et al.* 2005, pag. 60) that the optimal IS probability measure $\tilde{P}$ should be "as similar as possible" to the original probability measure conditioned on the event of interest, which we define as $P^{(c)}$. How can we measure the discrepancy? The most commonly used approach consists in

using the Cross-Entropy or *Kullback-Leibler distance* (Kullback 1968) between the two distributions:

$$D(P^{(c)}, \tilde{P}) = E^{(c)} \log \frac{P^{(c)}}{\tilde{P}}. \tag{13}$$

As pointed out in the preceding section, minimization of the variance of the estimator may not be straightforward. On the other hand, the CE approach is usually easier to implement and an instrumental distribution that is good by the CE criterion tends to be good in terms of variance as well. In particular, the two methods give the same IS density if the minimization of the Kullback-Leibler distance is performed over all densities (Rubinstein and Kroese 2004, pag. 67).

In the present setup, the probability measure $\tilde{P}$ is restricted to be of the form (8) and is therefore identified by parameters π and t. When the probability measure $\tilde{P}$ is absolutely continuous with density $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is a vector of parameters, it can be shown that minimizing $D(P^{(c)}, \tilde{P})$ is equivalent to

$$\max_{\boldsymbol{\theta}} E^{(c)} \left(\sum_{i=1}^{k} \log(f_{\boldsymbol{\theta}}(X_i)) \right).$$

It is now clear that there is a close relation between entropy minimization (namely, minimization of (13)) and likelihood maximization. The log-likelihood of k observations is indeed given by

$$\sum_{i=1}^{k} \log(f_{\boldsymbol{\theta}}(x_i)) = k \int \log(f_{\boldsymbol{\theta}}(x)) P_k(dx) = -kD(P_k, P_{\boldsymbol{\theta}}) + \text{const}, \tag{14}$$

where P_k is the empirical distribution. Comparing (13) to (14), it follows that maximum likelihood results can be translated into minimum CE results by replacing P_k with $P^{(c)}$.

The second fundamental result we need in the following is borrowed from Asmussen (2000, lemma 5.6), to which the interested reader is referred for details and a proof. The lemma states that, given $A(c) = \{X_1 + \cdots + X_k > c\}$, where X_i's are subexponential distributions with distribution function F, $A(c)$ occurs if $k-1$ of the X_i's have distribution F and one has the conditional distribution of X given $X > c$. Notice that this is very similar to the definition of subexponential random variables.

How do these results combine to provide a method for determining the optimal parameters of the IS density? We explain this issue by focusing on our setup. First, notice that MLE of the parameters of a lognormal mixture is equivalent to MLE of the parameters of a normal mixture. To see why, recall that a normal mixture in two populations $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ is obtained by sampling from X_1 with probability π and from X_2 with probability $1-\pi$. On the other hand, a lognormal mixture in two populations $Y_1 \sim \text{Logn}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \text{Logn}(\mu_2, \sigma_2^2)$ is obtained by sampling from $Y_1 = e^{X_1}$ with probability π and from $Y_2 = e^{X_2}$ with probability $1-\pi$. Thus, when taking logarithms of Y_1 and Y_2, one samples from $\log(Y_1) = X_1$ with probability π and from $\log(Y_2) = X_2$ with probability $1-\pi$.

Therefore we can consider the MLEs of the parameters π and $\mu_2 \stackrel{\text{def}}{=} \mu + t$ of a normal mixture (see,

for example, Flury 1997, sect. 9.2):

$$\hat{\pi} = \frac{1}{k}\sum_{i=1}^{k}\pi_{1i} = \int_{-\infty}^{\infty}\pi_{1i}F_k(dx), \tag{15}$$

$$\hat{\mu}_2 = \frac{1}{k(1-\hat{\pi})}\sum_{i=1}^{k}\pi_{2i}x_i = \frac{1}{1-\hat{\pi}}\int_{-\infty}^{\infty}x\pi_{2i}F_k(dx), \tag{16}$$

where F_k is the empirical distribution function and $\pi_{2x} = 1 - \pi_{1x}$ is the posterior probability that x belongs to the second population. Defining ϕ_{μ,σ^2} as the $N(\mu,\sigma^2)$ density, π_{1x} is given by

$$\pi_{1x} = \frac{\pi\phi_{\mu,\sigma^2}(x)}{\pi\phi_{\mu,\sigma^2}(x) + (1-\pi)\phi_{\mu_2,\sigma^2}(x)}. \tag{17}$$

Putting $\tilde{c} = \log(c)$, the conditional density of $X|X > \tilde{c}$ is equal to

$$\phi^{(\tilde{c})}_{\mu,\sigma^2}(x) = \frac{\phi_{\mu,\sigma^2}(x)}{1-\Phi_{\mu,\sigma^2}(\tilde{c})}\mathbf{1}_{\{x>\tilde{c}\}},$$

where Φ_{μ,σ^2} is the $N(\mu,\sigma^2)$ distribution function. Now the values of π and μ_2 (π^* and μ_2^*, say) which minimize entropy are given by (15) and (16) with F_k replaced by $\Phi^{(\tilde{c})}_{\mu,\sigma^2}$. More precisely, from Asmussen (2000, lemma 5.6) we know that $(k-1)$ observations have density ϕ_{μ,σ^2} and one has density $\phi^{(\tilde{c})}_{\mu,\sigma^2}$, so that

$$\pi = \frac{k-1}{k}\int_{-\infty}^{\infty}\pi_{1x}\phi_{\mu,\sigma^2}(x)dx + \frac{1}{k}\cdot\frac{1}{1-\Phi_{\mu,\sigma^2}(\tilde{c})}\int_{\tilde{c}}^{\infty}\pi_{1x}\phi_{\mu,\sigma^2}(x)dx, \tag{18}$$

$$\mu_2 = \frac{k-1}{k(1-\pi)}\int_{-\infty}^{\infty}x\pi_{2x}\phi_{\mu,\sigma^2}(x)dx + \frac{1}{k(1-\pi)}\cdot\frac{1}{1-\Phi_{\mu,\sigma^2}(\tilde{c})}\int_{\tilde{c}}^{\infty}x\pi_{2x}\phi_{\mu,\sigma^2}(x)dx. \tag{19}$$

Now the key to the solution of the system formed by (18) and (19) consists in noting that equations (17), (18) and (19) are the equations of the EM algorithm for maximum likelihood estimation of the parameters of a random variable X distributed as a two-population normal mixture with parameters (μ,σ^2) and (μ_2,σ^2) respectively, where $(k-1)$ observations are from the mixture itself and one observation is from the distribution of $X|X \geq \tilde{c}$. In particular, (17) implements the E-step, (18) and (19) the M-step. Hence, in order to get the optimal values π^* and μ_2^*, we iterate (17), (18) and (19) until convergence; note that the integrals in (18) and (19) have to be solved numerically at each iteration.

3.2 The standard Cross-Entropy approach

In this subsection we sketch the standard CE approach. By "standard CE approach" (Rubinstein and Kroese 2004, sect. 2.3) we mean that the IS density is chosen in the same parametric family of the variable of interest and that the CE method is used to find the optimal tilting parameter. In the present setup this implies $g \sim \text{Logn}(\mu_2,\sigma^2)$, where $\mu_2 = \mu + t$ $(t \geq \mu)$. Therefore, the optimal value of t is determined by minimizing the Kullback-Leibler distance, and in this case the analogy with maximum likelihood estimation provides a simple solution: given k observations $x_1,\ldots,x_k$, the MLE of μ is given by:

$$\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k}x_i = \int_{-\infty}^{\infty}xF_k(dx),$$

where $x_i = \log(y_i)$. Following the same reasoning of the preceding subsection and using the well-known expression of the expected value of the truncated normal distribution, the optimal value of μ_2 can be obtained analytically:

$$\mu_2^* = \frac{k-1}{k}\int_{-\infty}^{\infty} x\phi(x)dx + \frac{1}{k}\cdot\frac{1}{1-\Phi(c)}\int_{(\tilde{c}-\mu)/\sigma}^{\infty} x\phi(x) = \tag{20}$$
$$= \frac{k-1}{k}\mu + \frac{1}{k}\left(\mu + \sigma\frac{\phi((\tilde{c}-\mu)/\sigma)}{1-\Phi((\tilde{c}-\mu)/\sigma)}\right).$$

Notice that the optimal value μ_2^* completely determines the IS density.

3.3 The Adaptive Cross-Entropy approach

As pointed out by Rubinstein and Kroese (2004, p. 38), the standard CE approach (from now on ST) does not work well if the probability of the event of interest is too small (below 10^{-5}); to overcome this difficulty, Rubinstein and Kroese (2004, sect. 3.4) propose a multilevel algorithm which is based on a two-step adaptive procedure (also called "Adaptive Cross-Entropy" - ACE) where not only the parameter μ but also the threshold c is updated at each iteration. While referring the interested reader to Rubinstein and Kroese (2004) for theoretical properties of this technique, in the next section we give some details about the implementation to our setup. It is worth stressing that the ACE approach is computationally quite heavy, because each iteration consists of two phases: in the present setup, the updating of μ is done using standard numerical methods, while c is updated by means of MC simulation.

4 Some simulation experiments

In this section we focus on the estimation of tail probabilities for the random variable $Y = \sum_{i=1}^{k} X_i$ with $X_i \sim \text{Logn}(0,1)$ and $k = 10$. Table 1 shows the optimal values of the parameters of the IS density in the three approaches (DM, ST and ACE) presented in the preceding section. In the first case the IS density is a defensive mixture and the optimal parameters π^* and t_{DM}^* are determined by means of the EM algorithm. In the remaining two cases the IS density is a lognormal density and the optimal value of t is found respectively using (20) and the multilevel algorithm proposed by Rubinstein and Kroese (2004, sect. 3.4). The latter algorithm is implemented with $N = 10000$ MC replications and a sample quantile equal to $1-\rho = 0.99$; different values of ρ produce almost identical results. According to remark 3.9 in Rubinstein and Kroese (2004, p. 74), we iterate the algorithm ten more times after the stopping criterion is satisfied.

TABLE 1 HERE

From the table it can be seen that in the DM approach the optimal value of π remains almost constant as c increases, but the tail of the IS density gets heavier because t_{DM} increases. As for the other two approaches, in ACE the parameter t is much more sensitive to c than in ST.

As pointed out above, the performance of any IS estimator is related to the features of the distribution of $r_k(\boldsymbol{X})$. Table 2 gives some details about this distribution in the DM, ST and ACE approaches with $c \in \{65, 80, 100, 150, 200, 300, 400, 500\}$ and $k = 10$.

TABLE 2 HERE

The results show quite clearly that $r_{k,DM}$ has many more desirable distributional properties than $r_{k,ST}$ and $r_{k,ACE}$. In particular, we stress that approximately half of the observations simulated in the DM approach are used for the computation of the estimator, even for the largest thresholds, and this percentage is much larger than for ST and ACE (see the values of f^c in table 2). Moreover, it is worth pointing out that the distribution of $r_{k,DM}$ maintains satisfactory features as c increases. On the other hand, in the ST approach the average of the weights remains approximately stable as c grows, but the number of observations exceeding the threshold is very small, especially for large c. Finally, and somewhat surprisingly, we see that $r_{k,ACE}$ has quite poor properties as well: the average of the weights tends to zero as c increases, and the number of observations actually used for the computation of the tail probability reduces considerably as c gets large. These features are typical of the weight degeneracy mentioned in Section 2.4. The different properties of the distributions of $r_{k,DM}$ and $r_{k,ACE}$ are clearly related to the fact that the weights in the DM case are bounded above, whereas in the ACE approach, they are not. As we are now going to see, these results have a strong impact on the properties of the resulting estimator.

In the same setup used above for the simulation of $r_k(\boldsymbol{X})$, we investigate the properties of the three estimators of p_k with $N = 10,000$ (see Table 3). To assess the stability of the estimators, we repeated the simulation $B = 1,000$ times and computed the MC estimate of the standard error of the estimator $\hat{\text{se}}(\hat{p}) = (1/B)\text{var}(p^{(i)})$, where $p^{(i)}$ $(i = 1, \ldots, B)$ is the estimate obtained at the i-th replication. The estimators are not guaranteed to be unbiased; therefore, neither the standard deviation nor the relative error are good measures of performance, because both of them are only appropriate when the estimator is unbiased. As a consequence we estimated a version of the relative error (we call it "MSE Relative Error") based on the MSE instead of the variance: $\tau_{MSE} \stackrel{\text{def}}{=} \sqrt{\text{MSE}(\mathbf{1}_{\{X_i \geq c\}})}/E(\mathbf{1}_{\{X_i \geq c\}})$. The problem with τ_{MSE} is that we do not know the true value of p and therefore, in principle, we can't compute the MSE; however, using here the conclusions drawn from the results in Figure 3 (see below), $\hat{p}_{DM}$ seems to have essentially reached convergence for sample sizes larger than $100,000$. As the estimator is consistent, we computed $\hat{p}_{DM}$ for all c's with a sample size as large as $N = 5,000,000$ and computed τ_{MSE} treating the value so obtained as the true value. Table 3 displays the results.

TABLE 3 HERE

Before commenting the outcomes in table 3, we investigate the asymptotic properties of the estimators and give some graphical representations. Figure 2 shows the histograms of the simulated distributions of $\hat{p}_{DM}$ and $\hat{p}_{ACE}$ respectively for the cases $c = 65$ and $c = 500$; whereas the first estimator is approximately normal as expected from the theory of MC simulation, the latter has a very skewed distribution, in particular for $c = 500$.

FIGURE 2 HERE

Figure 3 shows the convergence of $\hat{p}_{DM}$ and $\hat{p}_{ACE}$ as a function of N for two different values of c, namely $c = 200$ and $c = 500$; notice also that the final sample size is $N_{ACE} = 10,000,000$ for ACE and only $N_{DM} = 500,000$ for DM, because $\hat{p}_{DM}$ remains essentially constant for $N_{DM} > 500,000$.

FIGURE 3 HERE

As expected, the DM approach clearly outperforms ACE, but there is a marked difference between the two cases: for $c = 200$, $\hat{p}_{ACE}$ shows an acceptable precision, although only for the largest sample sizes (the final values are $\hat{p}_{DM} = 9.1 \cdot 10^{-7}$ and $\hat{p}_{ACE} = 3.75 \cdot 10^{-7}$). On the other hand, for $c = 500$ $\hat{p}_{ACE}$ does not seem to reach convergence, even though it moves in the "right" direction as the sample size increases.

From Table 3 and Figure 3 we can derive some interesting conclusions. First, the DM approach is always preferable to ST and ACE, because it is approximately unbiased even for small sample sizes and has a much lower MSE Relative Error. Notice that the value of τ_{MSE} for $\hat{p}_{DM}$ remains stable up to the second decimal digit for all values of c, whereas τ_{MSE} deteriorates considerably for $\hat{p}_{ACE}$ when the probability of interest gets smaller: in particular, for $c = 500$ the ratio of the two MSE Relative Errors is approximately equal to 25.

The estimator $\hat{p}_{ACE}$ is approximately unbiased and has a low MSE Relative Error only for very large N. This fact has the obvious implication that computing time increases; moreover, it is not clear how large N should be. As for the ST approach, it only performs well for relatively large probabilities; for the largest values of c the tilting parameter is clearly too small. For these reasons, the DM approach seems to be preferable in all instances.

Finally, it may be relevant to measure the performance of the DM approach when the parameters are not optimized. This issue is of interest because, as pointed out by Hesterberg (1985), the results may not be very dependent on the exact shape of the second component of the mixture. If this is the case, it means that the overall cost of sampling is just slightly larger with non-optimized parameters, so that one may consider avoiding optimization (which has its own costs, both in human time and in machine time) and use parameters chosen in a more heuristic way (see Hesterberg 1995, sect. 6, for some possible solutions). Table 4 is similar to Table 3, but the results only concern the DM approach and are obtained with $\pi = 0.9$ and $t_{DM} \in \{3, 4, 5, 6, 7, 8\}$.

TABLE 4 HERE

The outcomes are quite interesting. First, it can be seen that the estimator is essentially unbiased for all the values of t_{DM} used in the experiment. However, the stability of the estimator is clearly dependent on t_{DM}, and even a small departure from the optimal value causes a non-negligible increase of τ_{MSE}. Consider, for example, the case $c = 65$: the optimal τ_{MSE} is equal to 0.04 (see Table 3) and corresponds to $t_{DM} = 4.265$ (Table 1). Using $t_{DM} = 4$, which is not very different from the optimal value, gives a four times larger MSE Relative Error; with $t_{DM} = 6$, the MSE Relative Error is almost 40 times larger.

As for the cost of optimizing, DM has a much more favorable performance, because its computing time is approximately equal to 0.03 seconds for any c, whereas for ACE the time increases from 1.6 to 10.5 seconds as c gets large.

5 Computing tail probabilities in Operational Risk

Operational risk management (see Davis 2006 for an overview of problems and techniques) is usually defined as the area of risk management concerned with non-financial losses: it includes internal and external frauds, employment practices and workplace safety, clients, products, and business practices, damage to physical assets, business disruption and system failures, execution, delivery and process management. It has recently become more and more important, both because of the regulators' pressure and of the amount of losses.

Operational risk presents peculiar features with respect to market and credit risk; it follows that its measurement and management require different tools. In particular, the distribution of losses is mostly modeled directly because operational losses are not related to underlying financial factors. This characteristic has been the key to the development of a purely statistical approach which assumes a fully parametric model for the losses and estimates its parameters using historical observations.

The *Loss Distribution Approach* is the most advanced approach contemplated by the Basel II Accord; it is based on the well known actuarial methodology which estimates the whole loss distribution for each business line by modeling separately the frequency and the severity of losses; see Embrechts *et al.* (1997) or Klugman *et al.* (1998) for details. The standard parametric model currently used in applications is the compound Poisson-lognormal distribution. Thus, the joint probability density function of $Y = \sum_{i=1}^{K} X_i$ and K over a fixed time horizon T is given by:

$$f_{Y,K}(y,k) = P(K=k) \cdot f_Y(y),$$

where $K \sim \text{Poisson}(\lambda)$ and $X_i \sim \text{Logn}(\mu, \sigma^2)$ model respectively the frequency and severity of losses. The marginal distribution of Y is the infinite mixture (often called compound) distribution

$$f_Y(y) = \sum_{i=0}^{\infty} P(K=i) \cdot f_{Y_i}(y). \tag{21}$$

The most common risk measure is the Value at Risk (VaR); the VaR at level α is the α quantile of (21), so that $1-\alpha$ is the tail probability of L. However, (21) is not known in closed form; therefore, the only way of estimating quantiles relies on MC simulation. Moreover, the Basel II Accord prescribes large confidence levels (up to 99.9%), and CMC encounters the problems mentioned in the preceding sections.

The CMC procedure consists of the following steps:

1. simulate a random number k^* from the Poisson(λ) distribution;
2. simulate k^* random numbers $x_1^*, \ldots, x_{k^*}^*$ from the Logn(μ, σ^2) distribution and compute $y^* = \sum_{i=1}^{k^*} x_i^*$.

Repeating B times (where B is a large number) steps 1. and 2. above, we simulate the loss distribution; the VaR at confidence level α is given by the α quantile of the empirical distribution.

As for the application of IS to the compound Poisson distribution, some methods are discussed by Asmussen and Glymm (2007, chap. 6) for the light-tailed case; in the heavy-tailed setup, Asmussen and Kroese (2006) propose a solution which also incorporates control variates and stratification techniques. Here we show how the procedure outlined in this paper can be extended to the random sum framework. Recall the functional form of (21) and put $P(K = k) = q_k$, $k = 0, 1, \ldots$. We have

$$P(Y > c) = \sum_{i=1}^{\infty} q_i \int_c^{\infty} f_{Y_i}(y)dy; \tag{22}$$

thus, we can apply to each summand the IS procedure developed above. The only problem is that we have to truncate the series; however, given the properties of the Poisson distribution, the series can usually be truncated after few terms. Obviously, the decision has to be made on the basis of the value of λ; for example, with $\lambda = 1$, $P(K \geq 7) \approx 8.3 \cdot 10^{-5}$; this implies that, if we simulate 10000 random numbers to estimate (22), we can stop when $K = 6$ or at most when $K = 7$, so that the sum contains just 5 (respectively 6) summands (the first term, corresponding to $k = 0$, can also be discarded because the loss is zero).

In this application we consider the example used by Bee (2006) for estimating a Poisson-lognormal model with truncated data; the estimated parameter values of the Poisson-lognormal model were $\hat{\lambda} = 6.931$, $\hat{\mu} = 1.404$, $\hat{\sigma}^2 = 2.823$ and $N_T = 950.63$ (N_T is the estimated number of truncated data; see Bee 2006 for details). First of all, we have to compute the optimal parameters of the IS densities. We do not give the numerical values of the optimal parameters for all c's and k's but consider that, for example, for $c = 1000$ in the DM approach π^* increases from 0.467 when $k = 2$ to 0.935 when $k = 20$ (as seen in sect. 3, the case $k = 1$ is special because $r_1(x)$ is bounded even for $\pi = 0$, so that there is no need to use the DM approach: for $k = 1$ we obtained $\pi \approx 6^{-10}$). As for t_{DM}, it decreases from 7.35 for $k = 1$ to 6.76 for $k = 20$ (with $\lambda = 6.931$, the probability of a value larger than 20 is approximately equal to $1.25 \cdot 10^{-5}$; hence, we truncated the series (22) at $k = 20$). In the ST approach t_{CE} ranges from 5.95 for $k = 1$ to 0.297 when $k = 20$, and in the ACE approach from 5.93 to 0.390.

Table 5 shows estimated tail probabilities and MSE relative errors obtained with the DM, ST, ACE and CMC approaches with $N = 10000$; not surprisingly, and according to the results of the simulations of the preceding section, DM has the best performance; notice, however, that the ST approach also works well if the threshold is not too large.

TABLE 5 HERE

6 Conclusions

Estimating rare events probabilities by means of the standard MC method is in general very inefficient. On the other hand, when the functional form of the density of the variable of interest is not known,

deterministic numerical approaches cannot be applied. Borrowing an idea first introduced by Hesterberg (1995), in this paper we have developed a mixture-based IS strategy for the estimation of tail events probabilities when the distribution of interest is a finite sum of lognormal random variables and for the compound Poisson-lognormal distribution. We solved the problem of finding the values of the parameters of the mixture by means of the CE method: in particular, by exploiting the relationship between minimal CE and maximum likelihood, we showed that the parameters can be found using the EM algorithm. With the help of simulation experiments we verified that this technique works better than the standard and adaptive CE approaches. Finally, we applied the methodology to the computation of tail probabilities in operational risk.

Several issues remain open to future research. First, although the lognormal distribution is by far the most common choice, in practice different distributional hypotheses for the severity of losses are sometimes used, according to the estimated tail heaviness: in particular the Gamma or the Generalized Pareto. While in the first case the implementation of IS should not be too difficult, as the Gamma distribution has the moment generating function, so that we can use the approach based on tilted densities, in the latter setup the problem would require some further research. Second, in the Poisson-lognormal compound distribution, importance sampling could also be applied by leaving unchanged the parameters of the lognormal and twisting the Poisson parameter, i.e. increasing the number of losses. Although this topic is aside from the main object of interest of this paper, it should be studied for its possible relevance in applications. Finally, it may be of interest to compute other risk measures: the so-called Expected Shortfall is the conditional expectation of the loss given that it exceeds some fixed value. The extension of the methods proposed here to the computation of Expected Shortfall requires further investigation.

References

Asmussen, S. (2000). *Ruin Probabilities*. London: World Scientific.

Asmussen, S. Binswanger, K. (1997). Simulation of ruin probabilities for subexponential claims. ASTIN Bulletin 27: 297-318.

Asmussen, S. Glymm, P.W. (2007). *Stochastic Simulation: Algorithms and Analysis*. New York: Springer.

Asmussen, S. Kroese, D.P. (2006). Improved algorithms for rare event simulation with heavy tails. Advances in Applied Probability 38: 545-558.

Asmussen, S. Binswanger, K. Højgaard, B. (2000). Rare events simulation for heavy-tailed distributions. Bernoulli 6: 303-322.

Asmussen, S. Kroese, D.P. Rubinstein, R.Y. (2005). Heavy tails, importance sampling and cross-entropy. Stochastic Models 21: 57-76.

Bee, M. (2006). Estimating the parameters in the loss distribution approach: how can we deal with truncated data? In: Davis E. ed., *The Advanced Measurement Approach to Operational Risk*. London: Risk Books.

Basel Committee on Banking Supervision (2005). *Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework*. www.bis.org.

Casella, G. Robert, C.P. (2004). *Monte Carlo Statistical Methods*, second edition. New York: Springer.

Davis, E. ed. (2006). *The Advanced Measurement Approach to Operational Risk*. London: Risk Books.

Davison, A.C. Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

Dempster, A.P. Laird, N.M. Rubin, D.B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm (with discussion). Journal of the Royal Statistical Society B 39: 1-38.

Durrett, R. (1996). *Probability: Theory and Examples*. Belmont: Duxbury Press.

Embrechts, P. Klüppelberg, C. Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. New York: Springer.

Flury, B. (1997). *A first course in multivariate statistics*. Springer: New York.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57: 1317-1340.

Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. New York: Springer.

Hammersley, J.M. Handscomb, D.C. (1964). *Monte Carlo Methods*. London: Methuen.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. Technometrics 37: 185-194.

Klugman, S.A. Panjer, H.H. Willmot, G.E. (1998). *Loss Models - From Data to Decisions*. New York: Wiley.

Kullback, S. (1968). *Information Theory and Statistics*. New York: Wiley.

Mikosch, T. Nagaev, A.V. (1998). Large deviations of heavy-tailed sums with applications in insurance. Extremes 1: 81-110.

Moran, P.A.P. (1984). *An Introduction to Probability Theory*. Oxford: Oxford Science Publications.

Ross, S.M., 2006. *Simulation*, fourth edition. San Diego: Academic Press.

Rubinstein, R.Y. (1997). Optimization of computer simulation models with rare events. European Journal of Operational Research 99: 89-112.

Rubinstein, R.Y. Kroese, D.P. (2004). *The Cross-Entropy Method*. New York: Springer.

Rubinstein, R.Y. Kroese, D.P. (2008). *Simulation and the Monte Carlo Method*, second edition. New York: Wiley.

Sadowsky, J.S. (1993). On the optimality and stability of exponential twisting in Monte Carlo simulation. IEEE Transactions on Information Theory IT-39: 119-128.

Table 1: Optimal parameter values in the three approaches with $k = 10$ and $X_i \sim \text{Logn}(0, 1)$.

	π^*	t^*_{DM}	t^*_{ST}	t^*_{ACE}
$c = 65$	0.893	4.265	0.439	0.636
$c = 80$	0.895	4.489	0.459	0.540
$c = 100$	0.896	4.726	0.481	1.251
$c = 150$	0.898	5.148	0.520	1.150
$c = 200$	0.899	5.441	0.548	2.043
$c = 300$	0.899	5.850	0.587	2.287
$c = 400$	0.900	6.137	0.615	2.455
$c = 500$	0.900	6.359	0.637	2.366

Table 2: Descriptive statistics for the distributions of $r_{k,DM}$, $r_{k,ST}$ and $r_{k,ACE}$ with $k = 10$; "max^c" is the maximum of the weights corresponding to observations such that $\sum_{i=1}^{k} x_i > c$, "$\bar{r}^c$" is the average of the observations such that $\sum_{i=1}^{k} x_i > c$, "f^c" is the ratio of the number of observations such that $\sum_{i=1}^{k} x_i > c$ to the number of simulations.

	$c = 65$	$c = 80$	$c = 100$	$c = 150$	$c = 200$	$c = 300$	$c = 400$	$c = 500$
$\bar{r}_{DM}$	0.94	0.99	0.87	0.96	1.02	1.02	1.09	0.94
$\bar{r}_{ST}$	1.05	1.02	1.03	1.39	0.98	0.87	0.84	0.81
$\bar{r}_{ACE}$	0.87	0.17	1.67	0.20	0.31	$9 \cdot 10^{-3}$	$4 \cdot 10^{-5}$	$7 \cdot 10^{-5}$
$\max_{DM}$	3.10	3.04	3.00	2.94	2.91	2.89	2.88	2.88
$\max_{ST}$	26.16	37.49	57.15	460.03	114.28	34.47	26.02	55.20
$\max_{ACE}$	52.78	22.96	1112.88	65.41	129.47	4.97	0.04	0.03
$\bar{r}^c_{DM}$	10^{-3}	$3 \cdot 10^{-4}$	10^{-4}	10^{-5}	$2 \cdot 10^{-4}$	$2 \cdot 10^{-7}$	$4 \cdot 10^{-8}$	$8 \cdot 10^{-9}$
$\bar{r}^c_{ST}$	0.15	0.07	0.01	-	-	-	-	-
$\bar{r}^c_{ACE}$	0.11	$4 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	$2 \cdot 10^{-9}$	$4 \cdot 10^{-10}$	$8 \cdot 10^{-14}$	$7 \cdot 10^{-7}$	$4 \cdot 10^{-16}$
$\max^c_{DM}$	0.03	$9 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$2 \cdot 10^{-4}$	$4 \cdot 10^{-5}$	$3 \cdot 10^{-6}$	$5 \cdot 10^{-7}$	10^{-7}
$\max^c_{ST}$	0.77	0.14	0.01	-	-	-	-	-
$\max^c_{ACE}$	1.47	0.03	$8 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-9}$	$9 \cdot 10^{-10}$	10^{-12}	$2 \cdot 10^{-15}$
f^c_{DM}	0.532	0.509	0.532	0.512	0.471	0.471	0.433	0.471
f^c_{ST}	0.007	0.004	0.002	0.001	0	0	0	0
f^c_{ACE}	0.017	0.115	0.047	0.016	0.004	0.004	0.033	0.005

Table 3: Some results from the simulation of the three estimators $\hat{p}_{DM}$, $\hat{p}_{ST}$ and $\hat{p}_{ACE}$ with $k = 10$.

	$c = 65$	$c = 80$	$c = 100$	$c = 150$
$\hat{p}_{DM}$	$5.71 \cdot 10^{-4}$	$1.74 \cdot 10^{-4}$	$4.89 \cdot 10^{-5}$	$4.84 \cdot 10^{-6}$
$\hat{p}_{ST}$	$5.71 \cdot 10^{-4}$	$1.75 \cdot 10^{-4}$	$4.93 \cdot 10^{-5}$	$4.86 \cdot 10^{-6}$
$\hat{p}_{ACE}$	$5.70 \cdot 10^{-4}$	$1.65 \cdot 10^{-4}$	$3.79 \cdot 10^{-5}$	$1.59 \cdot 10^{-6}$
$\tau_{MSE,\hat{p}_{DM}}$	0.04	0.04	0.04	0.04
$\tau_{MSE,\hat{p}_{ST}}$	0.30	0.57	0.95	3.30
$\tau_{MSE,\hat{p}_{ACE}}$	0.42	1.09	1.34	1.63
	$c = 200$	$c = 300$	$c = 400$	$c = 500$
$\hat{p}_{DM}$	$9.12 \cdot 10^{-7}$	$7.98 \cdot 10^{-8}$	$1.32 \cdot 10^{-8}$	$3.13 \cdot 10^{-9}$
$\hat{p}_{ST}$	$7.97 \cdot 10^{-7}$	$6.58 \cdot 10^{-8}$	$7.82 \cdot 10^{-10}$	$3.87 \cdot 10^{-8}$
$\hat{p}_{ACE}$	$5.39 \cdot 10^{-7}$	$2.95 \cdot 10^{-9}$	$9.77 \cdot 10^{-12}$	$9.37 \cdot 10^{-14}$
$\tau_{MSE,\hat{p}_{DM}}$	0.04	0.04	0.04	0.04
$\tau_{MSE,\hat{p}_{ST}}$	5.05	12.89	1.60	391.59
$\tau_{MSE,\hat{p}_{ACE}}$	7.26	1.22	1.00	1.00

Table 4: Some results concerning the estimation of $\hat{p}_{DM}$ with $k = 10$ with non-optimized parameters. The second subscript refers to the value of t_{DM} used.

	$c = 65$	$c = 80$	$c = 100$	$c = 150$
$\hat{p}_{DM,3}$	$5.71 \cdot 10^{-4}$	$1.74 \cdot 10^{-4}$	$4.91 \cdot 10^{-5}$	$4.83 \cdot 10^{-6}$
$\hat{p}_{DM,4}$	$5.71 \cdot 10^{-4}$	$1.75 \cdot 10^{-4}$	$4.90 \cdot 10^{-5}$	$4.88 \cdot 10^{-6}$
$\hat{p}_{DM,5}$	$5.68 \cdot 10^{-4}$	$1.75 \cdot 10^{-4}$	$4.89 \cdot 10^{-5}$	$4.90 \cdot 10^{-6}$
$\hat{p}_{DM,6}$	$5.78 \cdot 10^{-4}$	$1.74 \cdot 10^{-4}$	$4.86 \cdot 10^{-5}$	$4.91 \cdot 10^{-6}$
$\hat{p}_{DM,7}$	$5.96 \cdot 10^{-4}$	$1.57 \cdot 10^{-4}$	$4.65 \cdot 10^{-5}$	$4.82 \cdot 10^{-6}$
$\hat{p}_{DM,8}$	$4.41 \cdot 10^{-4}$	$1.70 \cdot 10^{-4}$	$4.15 \cdot 10^{-5}$	$5.15 \cdot 10^{-6}$
$\tau_{MSE,\hat{p}_{DM,3}}$	0.13	0.16	0.22	0.38
$\tau_{MSE,\hat{p}_{DM,4}}$	0.16	0.17	0.19	0.26
$\tau_{MSE,\hat{p}_{DM,5}}$	0.33	0.25	0.24	0.25
$\tau_{MSE,\hat{p}_{DM,6}}$	1.54	1.00	0.49	0.33
$\tau_{MSE,\hat{p}_{DM,7}}$	3.26	1.70	1.42	0.74
$\tau_{MSE,\hat{p}_{DM,8}}$	7.98	6.52	5.20	2.56
	$c = 200$	$c = 300$	$c = 400$	$c = 500$
$\hat{p}_{DM,3}$	$9.11 \cdot 10^{-7}$	$8.27 \cdot 10^{-8}$	$1.32 \cdot 10^{-8}$	$3.44 \cdot 10^{-9}$
$\hat{p}_{DM,4}$	$9.09 \cdot 10^{-7}$	$8.03 \cdot 10^{-8}$	$1.29 \cdot 10^{-8}$	$3.13 \cdot 10^{-9}$
$\hat{p}_{DM,5}$	$9.01 \cdot 10^{-7}$	$7.84 \cdot 10^{-8}$	$1.32 \cdot 10^{-8}$	$3.06 \cdot 10^{-9}$
$\hat{p}_{DM,6}$	$9.21 \cdot 10^{-7}$	$7.90 \cdot 10^{-8}$	$1.30 \cdot 10^{-8}$	$3.14 \cdot 10^{-9}$
$\hat{p}_{DM,7}$	$9.23 \cdot 10^{-7}$	$7.81 \cdot 10^{-8}$	$1.32 \cdot 10^{-8}$	$3.10 \cdot 10^{-9}$
$\hat{p}_{DM,8}$	$8.89 \cdot 10^{-7}$	$8.11 \cdot 10^{-8}$	$1.32 \cdot 10^{-8}$	$3.10 \cdot 10^{-9}$
$\tau_{MSE,\hat{p}_{DM,3}}$	0.56	1.13	1.73	2.74
$\tau_{MSE,\hat{p}_{DM,4}}$	0.33	0.50	0.63	0.87
$\tau_{MSE,\hat{p}_{DM,5}}$	0.25	0.29	0.36	0.40
$\tau_{MSE,\hat{p}_{DM,6}}$	0.30	0.27	0.28	0.30
$\tau_{MSE,\hat{p}_{DM,7}}$	0.54	0.40	0.36	0.33
$\tau_{MSE,\hat{p}_{DM,8}}$	1.63	1.01	0.75	0.60

Table 5: Estimates and MSE relative errors for operational risk data.

	$c = 1000€$	$c = 3000€$	$c = 5000€$	$c = 7500€$	$c = 10000€$
$\hat{p}_{DM}$	0.00486	0.00033	$8.53 \cdot 10^{-5}$	$2.77 \cdot 10^{-5}$	$1.22 \cdot 10^{-5}$
$\tau_{MSE}(\hat{p}_{DM})$	0.012	0.013	0.014	0.015	0.015
$\hat{p}_{ST}$	0.00486	0.00033	$8.40 \cdot 10^{-5}$	$2.74 \cdot 10^{-5}$	$1.18 \cdot 10^{-5}$
$\tau_{MSE}(\hat{p}_{ST})$	0.043	0.155	0.306	0.560	0.534
$\hat{p}_{ACE}$	0.00484	0.00031	$8.92 \cdot 10^{-5}$	$2.40 \cdot 10^{-5}$	$9.37 \cdot 10^{-6}$
$\tau_{MSE}(\hat{p}_{ACE})$	0.050	0.335	1.226	1.409	1.346
$\hat{p}^{MC}$	0.00484	0.00033	$7.66 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	$1.17 \cdot 10^{-5}$
$\tau_{MSE}(\hat{p}^{MC})$	0.137	0.517	1.041	2.058	2.969

Figure 1: MSE of the Monte Carlo estimator of E(X|X>c)

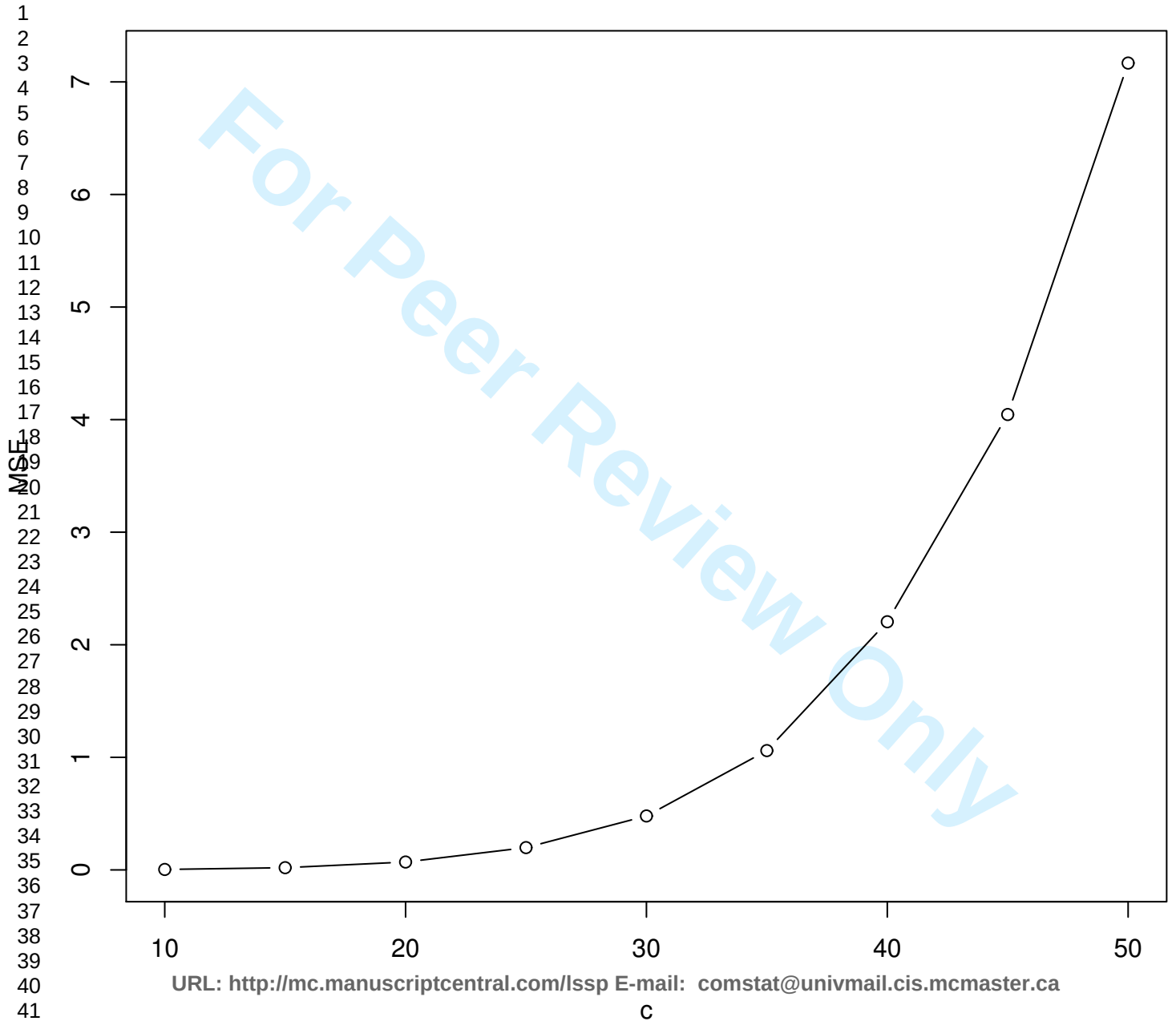


Fig. 2: Distribution of the DM and ACE estimators for c=200 and c=500

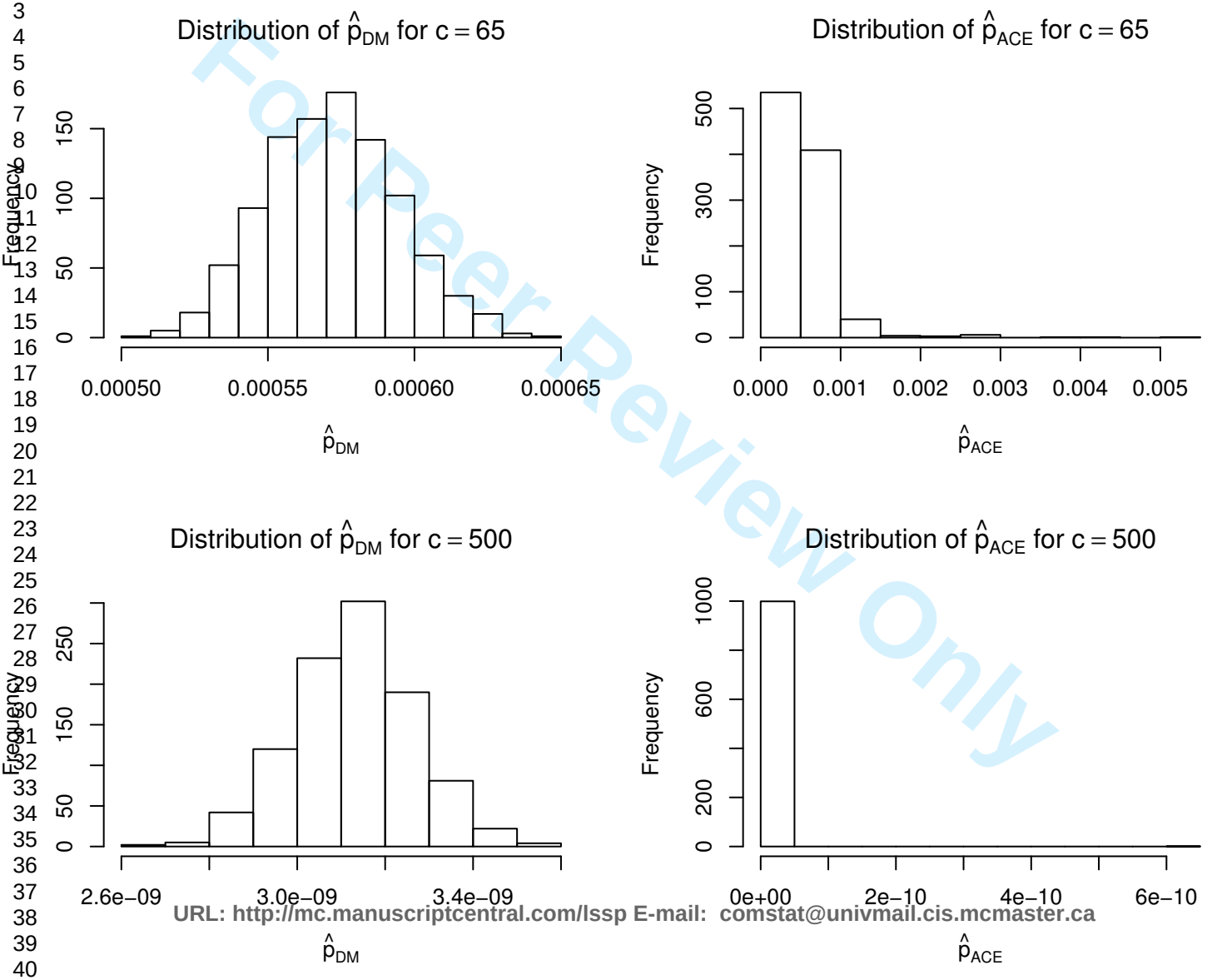
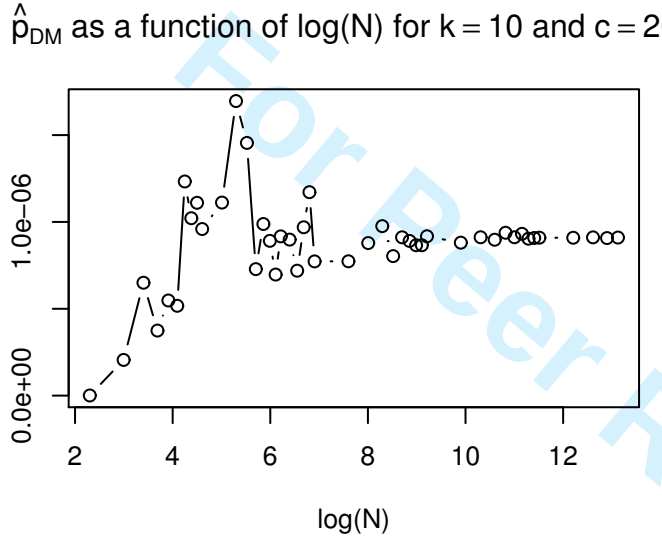
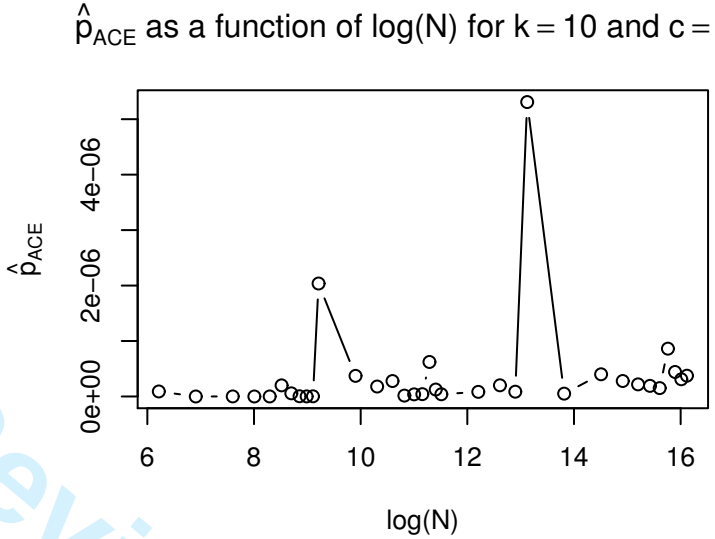
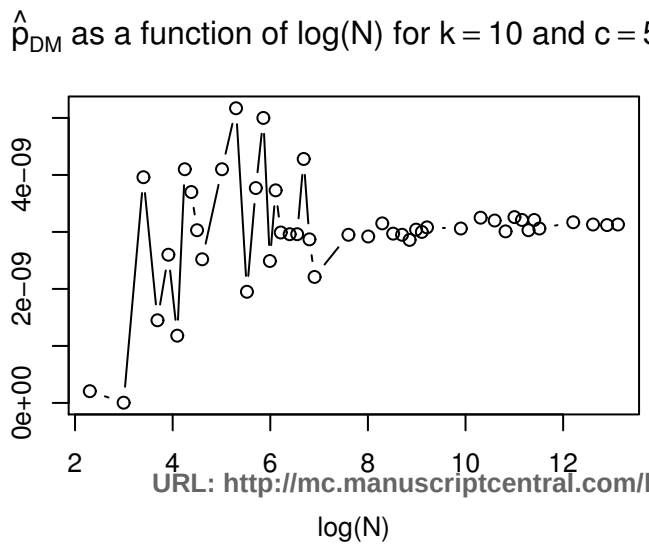
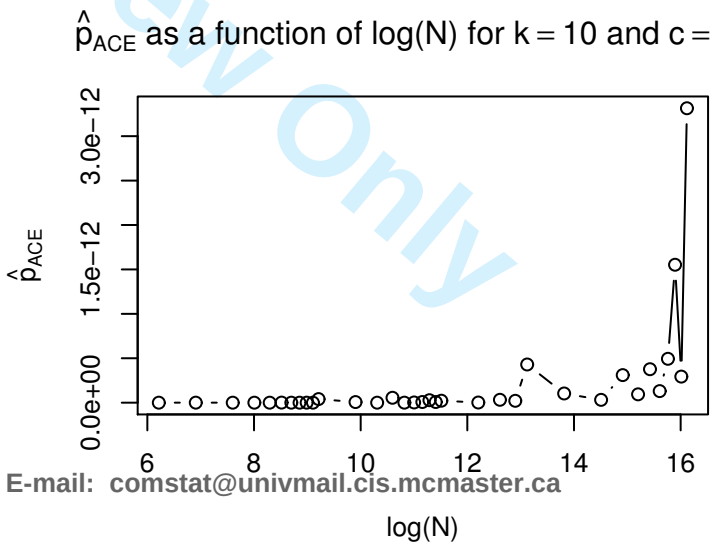


Fig. 3: Convergence of the DM and ACE estimators for c=200 and c=500









Responses to referees of the paper "Importance Sampling for Sums of Lognormal Distributions, with Applications to Operational Risk"

Referee 1

As a general comment, let me say that an earlier version of this paper contained a section devoted to IS for a single lognormal distribution. When writing the version revised by the referees, unfortunately, in several places I did not update the notation, so that, as pointed out by the referee (for example in comment 37), there are sentences referring explicitly or implicitly to the single lognormal case.

Specific Comments

As for comments 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 23, 26, 28, 29, 31, 32, 33, 35, 37, 38, 39, 40, 41, 43 and 46 the referee is right and I accepted (i.e., inserted in the paper) your remarks. I think the referee can just look at the new version of the paper to see whether now it is ok.

1. I ran the simulation again with $B = 10000$; in the first version I used a small B because, as the referee points out, in this case high accuracy is not necessary.

6. In the revised version only the MSE is shown, also according to the suggestion of another referee.

19, 20. Unfortunately, the two conditions were sloppily presented. I rewrote them completely.

21. I looked into Hammersley and Handscomb but could not find any reference to previous work on importance sampling. If the referee knows where any further reference on this can be found, I would be glad to include it in the paper.

22. One reason is that in section 5 k is the realization of a discrete r.v. K. I think in the actuarial and operational risk literature K is the common notation, so that I would rather continue to

use k. A second reason is that N is already used in equations (1) and (2) of example 1 with a different meaning.

24. Tilted densities are cited above (page 6^{14}) where there is also a reference to Ross (2006). There is no formal definition, if necessary I can add it.

25., 44. I added to the paper much of what you write. The confusion in the old version came from the fact that this part referred mainly to the case of a single lognormal distribution.

27., 45. I added some comments at the end of section 2 and the results of a simulation experiment in section 4.

30. As for the first part of the comment, e is the base of natural logarithm. As for the second part, the referee is right, the density must be strictly positive on $[0, a]$. As for the third part, I checked the article (Asmussen *et al.* 2000) where this density is proposed and could not find anything about this issue. As I do not use this approach in the paper, I decided not to explore it.

34. The sentence written in the old version of the paper was wrong, the optimal densities differ, this is actually clear from theorem 1 of the old version. Unfortunately, after a bit of reflection I conjectured that the methodology I use later for determining the IS density for the estimation of $P(Y > c)$ cannot be readily extended to the estimation of the conditional expectation. Thus, I decided to delete any reference to the latter problem, because, in case I am able to come up with a different solution, space constraints would not allow to add the details (both theoretical and numerical) to the paper. Given these remarks, I think that a solution to the problem of estimating the conditional expectation would be important enough for another paper to be written.

36. It is the form of the marginal density, and it is now written explicitly in the paper.

42. The theorem was presented without mentioning the hypothesis of independent marginals (both in f_k and in g_k). Having introduced this condition, I think it makes sense to leave it where it was in the first version.

45. This suggestion is interesting. I performed the comparison; the results are reported in table 4.

47. Concerning the CE approach, the optimization is done analytically, as already written just before formula (22). The ACE optimization is based on MC simulation, as written on page 13^{44-48}; however, in the new version I elaborate on this.

48. The referee is right, the notation is undefined. Considering that another referee complained about this, and suggested that reporting the lemma does not seem to be strictly necessary, I decided to remove it, and explain in words what it means.

Referee 2

In its present form, the manuscript is much too long.
I merged sections 2 and 3 of the old version into a single section. The paper was shortened in other places as well.

Conditions (6), (7) are very restrictive, and there are many excellent importance sampling algorithms where they do not hold.
I added this remark.

I can not make sense of the sentence "Notice that ..." p. 8.
The sentence written in the old version of the paper was wrong, the optimal densities differ, this is actually clear from theorem 1 of the old version. Unfortunately, after a bit of reflection I conjectured that the methodology I use later for determining the IS density for the estimation of $P(Y > c)$ cannot be readily extended to the estimation of the conditional expectation. Thus, I decided to delete any reference to the latter problem, because, in case I am able to come up with a different solution, space constraints would not allow to add the details (both theoretical and numerical) to the paper. Given these remarks, I think that a solution to the problem of estimating the conditional expectation would be important enough for another paper to be written.

I can not make sense of the statement some lines later that minimizing the variance is the same as minimizing the stated expected value.
This issue is thoroughly explained in the new version.

Simulation of compound Poisson sums is discussed, for example, in Asmussen & Glynn, Stochastic Simulation. Algorithms and Analysis. Springer 2007. There are much better methods than the truncation schemes discussed in the paper.
I added the reference. However, Asmussen and Glymm (2007) only discuss the light-tailed case; the heavy-tailed case is treated explicitly in Asmussen and Kroese (2006), which is now mentioned explicitly in section 5.

Rubinstein's 1981 has appeared in a second edition (co-authored with Kroese).
I updated the reference.

Referee 3

As for comments 2, 3, 4, 5, 6, 9, I accepted (i.e., inserted in the paper) the referee's remarks. I think the referee can just look at the new version of the paper to see whether now it is ok. I had a native English speaker read the paper, so that the language should now be correct.

1. As for the introduction, the referee is right, and I added more details about what I claim to be new. As for the abstract, I added something, but unfortunately space constraints are strict, and I couldn't add much.

7. I rewrote this part, it should now be clearer.

8. This section has been partly rewritten. The logarithmic transformation of the observations is now made explicit (the transformation $c^* = \log(c)$ was already written just above the place where it was first used). I fixed the references (I use Latex, but not Bibtex) and added an extra half space.

10. The referee is right, the notation is undefined. Considering that another referee complained about this, and that the lemma does not seem to be strictly necessary, I decided to remove it, and explain in words what it means.

11. I merged sections 2 and 3 of the old version into a single section, and the paper was shortened in other places according to the referee's suggestions.

12. I tried to prepare some 3D graphs, but they do not look very easy to read, so that in my opinion 2D graphs and tables are preferable (I decided to omit standard errors in table 3 because they are not of particular interest here). I added a table (number 4 in the new version) containing the results of a new analysis. Finally, I tried to improve the discussion of the results and the conclusions.

13. I enlarged the text of all the figures. As for figure 1, according also to a remark of another referee, I discarded the one showing the variance. In figure 3 the results are now shown as a function of the logarithm of the sample size.