AN M-ESTIMATOR FOR RARE-EVENT PROBABILITY ESTIMATION

Zdravko I. Botev  
School of Mathematics and Statistics 
The University of New South Wales 
Sydney, NSW 2052, AUSTRALIA

Ad Ridder  
Faculty of Economics and Business Administration 
Vrije University 
1081 HV Amsterdam, THE NETHERLANDS

ABSTRACT
We describe a maximum-likelihood type estimator, or M-estimator, for Monte Carlo estimation of rare-event probabilities. In this method, we first sample from the zero-variance measure using Markov Chain Monte Carlo (MCMC), and then given the simulated data, we compute a maximum-likelihood-type estimator. We show that the resulting M-estimator is consistent, and that it subsumes as a special case the well-known fixed-effort splitting estimator. We give a numerical example of estimating accurately the tail distribution of the sum of log-normal random variables under a Gaussian copula. The numerical results suggest that for this example the method is competitive.

1 INTRODUCTION
Suppose we wish to estimate a probability of the form
\[ \ell = \mathbb{P}(S(X) > \gamma), \quad X = (X_1, \ldots, X_d), \]
where: (a) \( S : \mathbb{R}^d \to \mathbb{R} \) is the so-called importance function; (b) \( X_1, \ldots, X_d \) are random variables with joint density \( f(x) \); and (c) \( \gamma \) is a threshold, which may be large enough to make \( \ell \) a rare-event probability. Such estimation problems arise in various contexts (Asmussen and Glynn 2007). For example, in financial engineering and under the Black-Scholes model, we may be interested in computing the tail distribution of the sum of dependent log-normal random variables under a Gaussian copula:
\[ S(X) = \exp(X_1) + \cdots + \exp(X_d), \]
where \( X \sim \mathcal{N}((\mu, \Sigma)), \) (Kortschak and Hashorva 2013, Asmussen et al. 2014, Laub et al. 2015).

Recently, a number of methods have been proposed for the estimation of \( \ell \) that use approximate simulation from the zero-variance measure via Markov chain Monte Carlo (Botev et al. 2011, Botev et al. 2013, Gudmundsson and Hult 2014, Botev et al. 2016). In this article we propose yet another such method. Similar to the existing approaches we first sample from the zero-variance measure via Markov Chain Monte Carlo. However, unlike existing methods, our approach then provides a maximum-likelihood-type estimator of the rare-event probability \( \ell \), given the simulated MCMC data. The proposed approach has been used in Bayesian statistics (Kong et al. 2003), but it has not been used in the rare-event simulation context (Huang and Botev 2013).

The proposed method has one main attraction compared to standard importance sampling. While in standard importance sampling the density has to satisfy a strict condition on its support and tail, in the M-estimation case, this restriction is much relaxed. As a result, at least in the examples we consider, it is simpler to incorporate analytical information, such as the asymptotic approximation of \( \ell \), into the estimation procedure.

We give a numerical example of estimating accurately the tail distribution of the sum of log-normal random variables under a Gaussian copula. Surprisingly the empirical results suggest that the proposed estimator may sometimes be more accurate than the corresponding tailor-made importance sampling estimator.
In addition, we show that this M-estimator is consistent under certain conditions, and that the well-known fixed-effort splitting estimator can be thought of as a special case of an M-estimator. This suggests a new approach to the theoretical analysis of such an estimator as future work.

2 M-ESTIMATOR OF RARE-EVENT PROBABILITY

To introduce the idea of a maximum-likelihood-type or M-estimator for the rare-event probability \( \ell \), it is convenient to think of \( \ell \) as a normalization constant \( \ell_s \) of the conditional density (here \( I\{\cdot\} \) is the indicator function of an event)

\[
\ell_s(x) = \frac{w_s(x)}{\ell_s} \overset{\text{def}}{=} f(x)I\{S(x) > \gamma\}.
\]

It is well known that the zero-variance importance sampling density for estimating \( \ell = \ell_s \) is the conditional pdf \( f_s \).

The typical importance sampling scheme proceeds as follows. Let \( f_1(x) = w_1(x)/\ell_1 \) be another reference or importance sampling density whose normalizing constant \( \ell_1 \) is known. Then, the standard importance sampling estimator with \( n \) samples is

\[
\ell_s^* = \frac{1}{n} \sum_{j=1}^{n} \frac{w_s(X_j)}{f_1(X_j)}, \quad X_1, \ldots, X_n \sim f_1.
\]

For acceptable performance, see (Kroese et al. 2011), we not only need the condition \( \{x : f_1(x) > 0\} \supseteq \{x : f_s(x) > 0\} \), but that the tails of \( f_1 \) to be at least as heavy as the tails of \( f_s \). Now suppose that we use \( f_s \) itself as part of a mixture importance sampling density that combines both \( f_1 \) and \( f_s \). In other words, we simulate \( n \) samples from the mixture density \( f \equiv \lambda_1 f_1 + \lambda_s f_s \), where \( \lambda_1 + \lambda_s = 1 \) are some weight fixed in advance. A plausible importance-sampling-type estimator then looks like:

\[
\hat{\ell}_s = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{w_s(X_j)}{f(X_j)} \right)^{1/\ell_s}, \quad X_1, \ldots, X_n \sim \hat{f}
\]

where the unknown normalizing constant \( \ell_s \) on the right is replaced with its estimator \( \hat{\ell}_s \). This substitution gives rise to a nonlinear equation for \( \hat{\ell}_s \). We thus define the M-estimator in this case with two-component mixture as the value \( \hat{\ell}_s \) that satisfies the nonlinear equation (2).

What have we gained by using \( \hat{\ell}_s \) as opposed to \( \ell_s^* \)? Compared with the traditional importance sampling estimator, \( \ell_s^* \), which requires the quite restrictive conditions on the tail and support of \( f_1 \) (\( \{x : f_1(x) > 0\} \supseteq \{x : f_s(x) > 0\} \) and \( \mathbb{E}_{f_1,f_s}(X)/f_1(X) < \infty \)), the tail and support restrictions on \( f_1 \) in the estimator \( \hat{\ell}_s \) are relaxed to the much weaker \( \{x : f_1(x) \times f_s(x) > 0\} \neq \emptyset \). In other words, the supports of \( f_1 \) and \( f_s \) need only overlap. For example, returning again to the log-normal probability \( \ell = \mathbb{P}(\exp(X_1) + \cdots + \exp(X_d) > \gamma) \), if \( f_1(x) \propto f(x)I\{\max_i \exp(X_i) > \gamma\} \), then this \( f_1 \) cannot be used in the standard importance sampling estimator \( \ell_s^* \), but it can be used in the estimator \( \hat{\ell}_s \), because the sets \( \{x : \exp(x_1) + \cdots + \exp(x_d) > \gamma\} \) and \( \{x : \max_i \exp(x_i) > \gamma\} \) overlap. The results in the numerical section suggest that for some problems (2) is a better estimator than some tailor-made schemes.

Before we proceed to show the consistency of \( \hat{\ell}_s \), we first generalize the method to a mixture of \( s \) pdfs, whose normalizing constants may or may not be known.

3 GENERALIZATION TO MULTIPLE COMPONENTS

Suppose we are given the sequence of densities (typically \( f_s \) being a zero-variance pdf)

\[
f_t(x) = \frac{w_t(x)}{\ell_t}, \quad t = 1, \ldots, s,
\]
where \( \ell_t \) are acting as normalizing constants. We assume that the supports of \( \{f_t\} \) satisfy Vardi’s connectivity condition (Gill et al. 1988, Vardi 1985). In other words, if we consider an undirected graph with \( s \) nodes and an edge between nodes \( i \) and \( j \) if and only if

\[
P(w_i(X) \times w_j(X) > 0) > 0, \tag{3}
\]

then the connectivity condition says that there exists a path between any two nodes of the graph. Recall that we are interested in estimating \( \ell_s \) and we assume that \( \ell_1 \) is known (we call \( f_1 \) a reference density), and without loss of generality, \( \ell_1 = 1 \). To estimate \( \ell_s \), we may simulate from each density as follows

\[
X_{t,1}, \ldots, X_{t,n_t} \sim f_t(x), \quad t = 1, \ldots, s,
\]

and collect the pooled sample as \( X_1, \ldots, X_n \), where the first \( n_1 \) samples are drawn from \( f_1 \), the next \( n_2 \) are drawn from \( f_2 \), and so on. Conceptually, this is not different from sampling \( n = n_1 + \cdots + n_s \) random variables with stratification from the mixture with \( s \) components

\[
f(x) = \frac{1}{n} \sum_{t=1}^s n_tf_t(x) = \sum_{t=1}^s \lambda_tf_t(x), \quad \lambda_t \overset{\text{def}}{=} n_t/n.
\]

We will henceforth assume that the proportions \( \lambda_t \) are fixed and do not change with the overall budget value \( n \). If we define the vector of parameters

\[
z = (z_1, \ldots, z_n)^\top \overset{\text{def}}{=} (-\log(1/\lambda_1), -\log(\ell_2/\lambda_2), \ldots, -\log(\ell_s/\lambda_s))^\top,
\]

it is clear that estimating \( \ell = (1, \ell_2, \ldots, \ell_s)^\top \), which includes the unknown \( \ell_s \), is equivalent to estimating \( z \). Then, we define the M-estimator of \( z \) as the solution of the optimization (Gill et al. 1988):

\[
\hat{z} = \arg\max_z \mathcal{D}_n(z),
\]

where we have the likelihood-lookalike objective function

\[
\mathcal{D}_n(z) \overset{\text{def}}{=} \frac{1}{n} \sum_{j=1}^n \log \left( \sum_{k=1}^s w_k(X_j) \exp(z_k) \right) - \sum_{k=1}^s \lambda_kz_k \tag{5}
\]

and

\[
\mathcal{D}(z) = \int \tilde{f}(x) \log \left( \sum_{k=1}^s w_k(x) \exp(z_k) \right) dx - \sum_{k=1}^s \lambda_kz_k.
\]

It may not be clear why the solution of this program yields a sensible estimator of the true \( z \) or \( \ell \). The first reason why this estimator makes sense is the following consistency result. (Here \( X_n \overset{\text{P}}{\rightarrow} X \) means that for any \( \epsilon, \delta \) pair we can find a large enough \( n \) so that \( \mathbb{P}(\|X_n - X\| > \epsilon) < \delta \), where \( \|\cdot\| \) is the Euclidean norm.)

**Proposition 1** (Consistency of estimator) If simulation from \( \tilde{f} \) (usually accomplished via MCMC) is such that a weak law of large numbers applies, \( \mathcal{D}_n(z) \overset{\text{P}}{\rightarrow} \mathcal{D}(z) \), uniformly in \( z \), then \( \hat{z} \overset{\text{P}}{\rightarrow} z \) as \( n \uparrow \infty \).

**Proof.** First, note that under the connectivity condition (3), Vardi et al. (Gill et al. 1988, Vardi 1985) show that \( \mathcal{D}_n \) is almost surely concave as \( n \uparrow \infty \). Next, we have by the assumption, \( \mathcal{D}_n(z) \overset{\text{P}}{\rightarrow} \mathcal{D}(z) \), that for any \( \tilde{z} \neq z \) (the first component is also \( \tilde{z}_1 = \hat{z}_1 = z_1 \)):

\[
\mathcal{D}_n(\tilde{z}) - \mathcal{D}_n(z) = \underbrace{\mathcal{D}_n(\tilde{z}) - \mathcal{D}(\tilde{z})}_{\overset{\text{P}}{\rightarrow} 0} + \underbrace{\mathcal{D}(\tilde{z}) - \mathcal{D}(z)}_{\overset{\text{P}}{\rightarrow} 0} + \underbrace{\mathcal{D}(\tilde{z}) - \mathcal{D}(z)}_{\overset{\text{P}}{\rightarrow} \alpha(\tilde{z}, z)},
\]
where we have denoted the last expression by
\[
\alpha(\tilde{z}, z) \overset{\text{def}}{=} \int \tilde{f}(x) \log \left( \sum_{k=1}^{s} \frac{w_k(x) \exp(\tilde{z}_k)}{\sum_{k=1}^{s} w_k(x) \exp(z_k)} \right) dx - \sum_{k=1}^{s} \lambda_k(\tilde{z}_k - z_k)
\]
Since \( \tilde{f}(x) = \sum_{t=1}^{s} \lambda_t f_t(x) = \sum_{t=1}^{s} \exp(z_t) w_t(x) \), we can apply Jensen’s inequality:
\[
\int \tilde{f}(x) \log \left( \sum_{k=1}^{s} \frac{w_k(x) \exp(\tilde{z}_k)}{\tilde{f}(x)} \right) dx = \int \tilde{f}(x) \log \left( \frac{\sum_{k=1}^{s} w_k(x) \exp(\tilde{z}_k)}{\tilde{f}(x)} \right) dx
\]
Jensen’s inequality \leq \log \left( \int \sum_{k=1}^{s} w_k(x) \exp(\tilde{z}_k) dx \right)
\leq \log \left( \sum_{k=1}^{s} \exp(\tilde{z}_k) \ell_k \right) = \log \left( \sum_{k=1}^{s} \lambda_k \exp(\tilde{z}_k - z_k) \right)
Therefore,
\[
D_n(\tilde{z}) - D_n(z) \overset{\mathbb{P}}{\to} \alpha(\tilde{z}, z) \leq \log \left( \sum_{k=1}^{s} \lambda_k \exp(\tilde{z}_k - z_k) \right) - \sum_{k=1}^{s} \lambda_k(\tilde{z}_k - z_k).
\]
Next, another application of Jensen’s inequality with the distribution \{\lambda_k\} yields
\[
\beta(\tilde{z}, z) \overset{\text{def}}{=} \log \left( \sum_{k=1}^{s} \lambda_k \exp(\tilde{z}_k - z_k) \right) - \sum_{k=1}^{s} \lambda_k(\tilde{z}_k - z_k) \geq 0.
\]
Since the logarithmic function is strictly concave, equality is achieved if and only if all the \( \exp(\tilde{z}_k - z_k) \) are equal. In other words, equality is achieved if and only if \( \tilde{z} = z + c \), where \( c \) is an arbitrary constant. In our case \( c = 0 \), because \( \tilde{z}_1 = z_1 \) by construction. In other words, \( \beta(\tilde{z}, z) > 0 \) for \( z \neq \tilde{z} \).
It follows that if \( z \neq \tilde{z} \), then with increasing probability as \( n \uparrow \infty \), \( D_n(\tilde{z}) - D_n(z) \) will be upper bounded by a strictly positive constant \( \beta \). That is, as \( n \uparrow \infty \) and \( z \neq \tilde{z} \) we have \( \mathbb{P}(D_n(z) > D_n(\tilde{z})) \uparrow 1 \).
Hence, using the fact that \( D_n(\tilde{z}) \leq D_n(z) \) for all \( \tilde{z} \), because \( \tilde{z} \) is, by construction, a global maximizer of the (almost surely) concave \( D_n \), we have as \( n \uparrow \infty \) and any \( \epsilon > 0 \),
\[
\mathbb{P}(\|\tilde{z} - z\| > \epsilon) = \mathbb{P}(\|\tilde{z} - z\| > \epsilon, D_n(z) \leq D_n(\tilde{z}), \text{for all } \tilde{z})
\leq \mathbb{P}(\|\tilde{z} - z\| > \epsilon, D_n(z) \leq D_n(\tilde{z}))
\leq \mathbb{P}(\tilde{z} \neq z, D_n(z) \leq D_n(\tilde{z}))
\leq 1 - \mathbb{P}(D_n(z) > D_n(\tilde{z}), \tilde{z} \neq z) \rightarrow 0
\]

A second way to see that estimator (4) is sensible is to solve for the gradient of \( D_n \) being zero: \( \nabla D_n(\tilde{z}) = 0 \). Rewriting this nonlinear system explicitly in \( \ell \) gives the system of moment-matching equations:
\[
\ell_t = \frac{1}{n} \sum_{j=1}^{n} w_t(X_j) \frac{\sum_{k=1}^{s} w_k(X_j) \lambda_k / \ell_k}{\sum_{k=1}^{s} w_k(X_j) \lambda_k / \ell_k}, \quad t = 2, \ldots, s,
\]
which appear to be the sample versions of the identities:
\[
\ell_t = \mathbb{E}_f \left[ \frac{\sum_{k=1}^{s} w_t(X) \lambda_k / \ell_k}{\sum_{k=1}^{s} w_t(X) \lambda_k / \ell_k} \right].
\]
\[362\]
Consider the special case in which all the densities \( f \) are of the form:

\[
f_t(x) = \frac{f(x) \mathbb{I}\{S(x) > \gamma_t\}}{\ell_t},
\]

where \(-\infty = \gamma_1 < \gamma_2 < \cdots < \gamma_s = \gamma\) (note that \(\gamma_s\) remains the same and indeed \(\ell_1 = 1\)). It is not difficult to see that these densities satisfy Vardi’s connectivity condition (3), because \(\mathbb{P}(S(X) > \gamma_j, S(X) > \gamma_i) = \mathbb{P}(S(X) > \max\{\gamma_i, \gamma_j\}) \geq \ell > 0\). Recall that simulation from \(\bar{f}\) with stratification is asymptotically equivalent to simulating \(n_1\) samples from \(f_1\), \(n_2\) samples from \(f_2\), and so on.

The next result shows that we can solve the system (6) exactly, obviating the need for Algorithm 1.

**Proposition 2** (Splitting and M-estimation) Let \(\mathbb{I}\) be an \(s \times n\) matrix with entries \(\mathbb{I}_{k,j} = \mathbb{I}\{S(X_j) > \gamma_k\}\). Then, the unique solution of (6) is:

\[
\hat{\ell}_t = \ell_1 \prod_{k=2}^{t} \frac{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k,j}}{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k-1,j}}, \quad t = 2, \ldots, s.
\]

**Proof.** We use induction. The formula is true for a matrix \(\mathbb{I}\) of size \(2 \times (n_1 + n_2)\), because the only solution to (6) is (dropping the hat accent from all \(\hat{\ell}\))

\[
\ell_2 = \ell_1 \frac{\sum_{j=1}^{n_1} \mathbb{I}_{2,j}}{\sum_{j=1}^{n_1} \mathbb{I}_{1,j}}.
\]

Now assume it is true for \(\mathbb{I}\) of size \((s-1) \times (n_1 + \cdots + n_{s-1})\), that is, we have

\[
\ell_t = \ell_1 \prod_{k=2}^{t} \frac{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k,j}}{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k-1,j}}.
\]
for $t = 2, \ldots, s - 1$. Expanding the size of the matrix to $s \times (n_1 + \cdots + n_s)$ gives that (6) is identical to

$$\ell_s = \sum_{j=1}^n \frac{\mathbb{I}_{k,j}}{n_k \ell_k} = \frac{\sum_{j=1}^n \mathbb{I}_{s,j}}{\sum_k n_k \ell_k},$$

from where we have $n_s + \ell_s \sum_{k \neq s} \frac{n_k}{\ell_k} = n_s + \sum_{j=1}^n \mathbb{I}_{s,j}$ and hence the only solution to (6) is:

$$\ell_s = \frac{\sum_{j=1}^n \mathbb{I}_{s,j}}{\sum_{k \neq s} \frac{n_k}{\ell_k}} = \ell_1 \prod_{k=2}^s \frac{\sum_{j=1}^n \mathbb{I}_{k,j}}{\sum_{j=1}^n \mathbb{I}_{k-1,j}},$$

because a direct calculation from left to right shows that

$$\frac{n_1}{\ell_1} + \frac{n_2}{\ell_2} + \cdots + \frac{n_{s-1}}{\ell_{s-1}} = \frac{n_1 \sum_{j=1}^{n_1+n_2} \mathbb{I}_{2,j}}{\ell_1} + \frac{n_3}{\ell_3} + \cdots + \frac{n_{s-1}}{\ell_{s-1}} = \frac{n_1 \sum_{j=1}^{n_1+n_2+n_3} \mathbb{I}_{3,j}}{\ell_1} + \cdots + \frac{n_{s-1}}{\ell_{s-1}}$$

$$= n_1 \prod_{k=2}^s \frac{\sum_{j=1}^{n_1+n_2+n_{k-1}} \mathbb{I}_{k,j}}{\ell_1} \prod_{k=2}^s \frac{\sum_{j=1}^{n_1+n_2+n_{k-1}} \mathbb{I}_{k-1,j}}{\ell_1} = n_1 \prod_{k=2}^s \frac{\sum_{j=1}^{n_1+n_2+n_{k-1}} \mathbb{I}_{k-1,j}}{\ell_1} \prod_{k=2}^s \frac{\sum_{j=1}^{n_1+n_2+n_{k-1}} \mathbb{I}_{k,j}}{\ell_1}$$

While not pursued here, this connection between splitting and M-estimation could be used to derive the asymptotic distribution of the splitting estimator using maximum likelihood results.

5 NUMERICAL ILLUSTRATION

Sums of Dependent Log-Normals. To illustrate the method, we consider the widely-studied log-normal tail distribution estimation mentioned in the introduction. We compare the accuracy of our M-estimator with that of the importance sampling vanishing error algorithm (abbreviated ISVE) proposed by (Asmussen et al. 2011). Recall that the probability of interest is $\ell = \ell_s = \mathbb{P}(\exp(X_1) + \cdots + \exp(X_d) \geq \gamma)$, which is the normalizing constant of the density ($s = 2$)

$$f_2(x) = f(x) \mathbb{I}_{\{S(x) \geq \gamma\}}$$

where $S(x) = \exp(x_1) + \cdots + \exp(x_d)$, and $f$ is the density of the multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma = (\Sigma_{ij})$.

We now estimate $\ell_2$ via the estimator (2) with $s = 2$ and reference density

$$f_1(x) = \frac{f(x) \sum_{j=1}^d \mathbb{I}_{\{\exp(x_i) > \gamma\}}}{\ell_1}, \quad \ell_1 \overset{\text{def}}{=} \sum_{j=1}^d \mathbb{P}(\exp(X_i) > \gamma), \quad (X_1, \ldots, X_d) \sim N(\mu, \Sigma).$$

In our numerical example we use the same parameters as (Asmussen et al. 2011), namely, a correlation coefficient

$$\frac{\Sigma_{i,j}}{\sqrt{\Sigma_{i,i} \Sigma_{j,j}}} = \rho, \quad \text{for all } i \neq j,$$
where the rest of the parameters of the density are set to: \(d = 10, \mu_i = i - 10, \sigma_i^2 = i\) for \(i = 1, \ldots, d\).

**Simulation from \(f_1\) and \(f_2\).** Sampling iid copies from the reference density is straightforward using the mixture representation

\[
f_1(x) = \sum_{j=1}^{d} \frac{\mathbb{P}(X_j > \log \gamma)}{\ell_1} f(x|X_j > \log \gamma).
\]

In other words, we select component \(J\) with probability \(\mathbb{P}(J = j) = \mathbb{P}(X_j > \log \gamma)/\ell_1\); then given \(J = j\), we sample from the truncated normal pdf \(f(x_j)\mathbb{I}(x_j > \log \gamma)/\mathbb{P}(X_j > \log \gamma)\); and finally we draw from the conditional multivariate normal \(f(x|X_j)\). Sampling (approximately) from \(f_2\) is accomplished using the Gibbs sampler (Gudmundsson and Hult 2014, Botev et al. 2016), whereby a single cycle of the Gibbs sampler consist of sequential draws \((i = 1, \ldots, d)\) from the truncated normal densities proportional to

\[
f(x_i|x_{-i}) \propto \mathbb{I}(x_i > \log \gamma - \sum_{j \neq i} \exp(x_j))
\]

We do not use any burn-in the Gibbs sampler and the initial state of the Markov chain is a sample simulated from \(f_1\).

Note that while there are no problems using the reference density \(f_1\) in our proposed M-estimator (2), it cannot be used as an importance sampling pdf in the standard importance sampling estimator (1) due to the lack of support on \(\{x : S(x) > \gamma, \max_i \exp(x_i) < \gamma\}\). This is one of the advantages of the M-estimator over the standard importance sampling estimator.

**Solving Nonlinear System.** To proceed with the implementation of the M-estimator, we simplify the moment-matching equation (2) as follows (note that \(S(X) > \gamma\) always under this simulation scheme):

\[
\hat{\ell}_2 = \frac{1}{n} \sum_{j=1}^{n} \frac{w_2(x_j)}{\lambda_1 w_1(x_j)/\ell_1 + \lambda_2 w_2(x_j)/\ell_2} = \frac{1}{n} \sum_{j=1}^{n} \frac{w_1(x_j) 1}{\ell_1 + \lambda_2/\ell_2} = \frac{p_0}{\ell_2} + \frac{p_1}{\ell_2} + \frac{p_2}{\ell_2} + \cdots + \frac{p_d}{\ell_2},
\]

where \(p_k\) is the number of \(X_j\)'s, which yield \(\sum_{i=1}^{d} \mathbb{I}(x_i > \log \gamma) = k\). Hence, our estimator \(\hat{\ell}_2\) solves the equation:

\[
p_0 + \frac{p_1}{\ell_2} + \frac{p_2}{\ell_2} + \cdots + \frac{p_d}{\ell_2} = n_2.
\]

Since this is not a splitting estimator, there is no simple analytical solution for the M-estimator and we have to resort to a numerical solution. Here we used \texttt{fzero.m} in Matlab to solve the equation without difficulty.

**Comparison between M-est. and ISVE.** Tables 1 and 2 show the estimates of \(\ell\) for various values of \(\gamma\) and \(\rho\). In all cases the algorithmic parameters are set to be \(\lambda_1 = \lambda_2 = 1/2\). Table 1 was created using \(n = 5 \times 10^5\) samples for both the M-estimator (M-est) and ISVE. Table 2 was created using \(n = 5 \times 10^6\) simulation runs. We did not use any burn-in for the Gibbs sampling. To estimate the relative variance of \(\hat{\ell}\) (which is asymptotically the variance of \(\hat{\ell}\)), we use the batch means method with 10 batches. In other words, the variance is estimated using 10 approximately independent sample averages of size \(n/10\).

The final column of both tables shows the estimate for the work normalized relative variance (WRNV), which takes into account the computational time for each method. The WRNV is defined as \(\tau \times \text{Var}(\hat{\ell})/\hat{\ell}^2\), where \(\tau\) is the CPU time.
Both M-estimator and ISVE use a total simulation effort of $\frac{\gamma \times 5}{10^5}$ of the reference pdf. Although, the reference value has the property that

$$-\ell \approx 0$$

for different values of $\gamma$. This is because the M-estimates (see Botev and Ridder). The asymptotic approximation here is $\ell \approx 1.7948 \times 10^{-5}$. Both M-estimator and ISVE use a total simulation effort of $n = n_1 + n_2 = 5 \times 10^5$.

Table 1: Empirical performance of M-estimator and ISVE for various values of the threshold parameter $\gamma = 5 \times 10^{c+3}$, $c = 1, \ldots, 14$ with $\rho = 0.999$. Both algorithms use a sample size of $n = n_1 + n_2 = 5 \times 10^5$.

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$\ell_1$</th>
<th>M-est.</th>
<th>ISVE estim.</th>
<th>M-est.</th>
<th>ISVE</th>
<th>M-est.</th>
<th>ISVE</th>
<th>WNRV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$5 \times 10^4$</td>
<td>0.000355</td>
<td>0.000409</td>
<td>0.000406</td>
<td>0.23</td>
<td>1.71</td>
<td>0.00044</td>
<td>15248</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^5$</td>
<td>1.794 \times 10^{-5}</td>
<td>2.212 \times 10^{-5}</td>
<td>2.177 \times 10^{-5}</td>
<td>0.23</td>
<td>3.09</td>
<td>0.00043</td>
<td>50267</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^6$</td>
<td>5.586 \times 10^{-7}</td>
<td>7.156 \times 10^{-7}</td>
<td>6.807 \times 10^{-7}</td>
<td>0.23</td>
<td>5.32</td>
<td>0.00042</td>
<td>1.4 \times 10^5</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^7$</td>
<td>1.057 \times 10^{-8}</td>
<td>1.384 \times 10^{-8}</td>
<td>1.444 \times 10^{-8}</td>
<td>0.23</td>
<td>11.74</td>
<td>0.00042</td>
<td>7.2 \times 10^5</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^8$</td>
<td>1.205 \times 10^{-10}</td>
<td>1.590 \times 10^{-10}</td>
<td>1.254 \times 10^{-10}</td>
<td>0.23</td>
<td>2.35</td>
<td>0.00042</td>
<td>29064</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^9$</td>
<td>8.230 \times 10^{-13}</td>
<td>1.086 \times 10^{-12}</td>
<td>3.781 \times 10^{-12}</td>
<td>0.23</td>
<td>76.90</td>
<td>0.00040</td>
<td>3.13 \times 10^7</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^{10}$</td>
<td>3.347 \times 10^{-15}</td>
<td>4.372 \times 10^{-15}</td>
<td>3.346 \times 10^{-15}</td>
<td>0.22</td>
<td>0.10</td>
<td>0.00040</td>
<td>56.12</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^{11}$</td>
<td>8.087 \times 10^{-18}</td>
<td>1.046 \times 10^{-17}</td>
<td>8.083 \times 10^{-18}</td>
<td>0.22</td>
<td>0.024</td>
<td>0.00039</td>
<td>2.99</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^{12}$</td>
<td>1.158 \times 10^{-20}</td>
<td>1.483 \times 10^{-20}</td>
<td>1.158 \times 10^{-20}</td>
<td>0.22</td>
<td>0.0018</td>
<td>0.00039</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^{13}$</td>
<td>9.827 \times 10^{-24}</td>
<td>1.245 \times 10^{-23}</td>
<td>1.641 \times 10^{-23}</td>
<td>0.22</td>
<td>40.12</td>
<td>0.00039</td>
<td>8.38 \times 10^6</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^{14}$</td>
<td>4.930 \times 10^{-27}</td>
<td>6.170 \times 10^{-27}</td>
<td>5.028 \times 10^{-27}</td>
<td>0.22</td>
<td>1.94</td>
<td>0.00039</td>
<td>19790</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^{15}$</td>
<td>1.462 \times 10^{-30}</td>
<td>1.804 \times 10^{-30}</td>
<td>1.462 \times 10^{-30}</td>
<td>0.22</td>
<td>0.00037</td>
<td>0.00038</td>
<td>0.00073</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^{16}$</td>
<td>2.562 \times 10^{-34}</td>
<td>3.123 \times 10^{-34}</td>
<td>2.563 \times 10^{-34}</td>
<td>0.22</td>
<td>0.00020</td>
<td>0.00038</td>
<td>0.00020</td>
<td></td>
</tr>
<tr>
<td>$5 \times 10^{17}$</td>
<td>2.651 \times 10^{-38}</td>
<td>3.198 \times 10^{-38}</td>
<td>2.652 \times 10^{-38}</td>
<td>0.22</td>
<td>0.000010</td>
<td>0.00037</td>
<td>5.21 \times 10^{-5}</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Empirical performance of M-estimator and ISVE for various values of the threshold parameter $\gamma = 5 \times 10^5$ with $\rho = 1 - 0.5^c$ and $c = 1, \ldots, 10$. The asymptotic approximation here is $\ell_1 \approx 1.7948 \times 10^{-5}$. Both M-estimator and ISVE use a total simulation effort of $n = n_1 + n_2 = 5 \times 10^6$.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>M-est.</th>
<th>ISVE estim.</th>
<th>M-est.</th>
<th>ISVE</th>
<th>M-est.</th>
<th>ISVE</th>
<th>WNRV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1 - 0.5^1$</td>
<td>1.8251 \times 10^{-5}</td>
<td>1.8212 \times 10^{-5}</td>
<td>0.063</td>
<td>0.14</td>
<td>0.00028</td>
<td>10270</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^2$</td>
<td>1.9336 \times 10^{-5}</td>
<td>1.9377 \times 10^{-5}</td>
<td>0.066</td>
<td>0.66</td>
<td>0.00031</td>
<td>208055</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^3$</td>
<td>2.0478 \times 10^{-5}</td>
<td>2.0355 \times 10^{-5}</td>
<td>0.069</td>
<td>0.91</td>
<td>0.00033</td>
<td>395812</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^4$</td>
<td>2.1246 \times 10^{-5}</td>
<td>2.1305 \times 10^{-5}</td>
<td>0.071</td>
<td>1.09</td>
<td>0.00035</td>
<td>566699</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^5$</td>
<td>2.1680 \times 10^{-5}</td>
<td>2.2332 \times 10^{-5}</td>
<td>0.072</td>
<td>1.37</td>
<td>0.00037</td>
<td>874993</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^6$</td>
<td>2.1928 \times 10^{-5}</td>
<td>2.2075 \times 10^{-5}</td>
<td>0.073</td>
<td>1.22</td>
<td>0.00037</td>
<td>707993</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^7$</td>
<td>2.2041 \times 10^{-5}</td>
<td>2.2091 \times 10^{-5}</td>
<td>0.073</td>
<td>1.26</td>
<td>0.00037</td>
<td>746153</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^8$</td>
<td>2.2099 \times 10^{-5}</td>
<td>2.2339 \times 10^{-5}</td>
<td>0.073</td>
<td>1.25</td>
<td>0.00037</td>
<td>733213</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^9$</td>
<td>2.2122 \times 10^{-5}</td>
<td>2.2208 \times 10^{-5}</td>
<td>0.073</td>
<td>1.30</td>
<td>0.00038</td>
<td>790156</td>
<td></td>
</tr>
<tr>
<td>$1 - 0.5^{10}$</td>
<td>2.2134 \times 10^{-5}</td>
<td>2.1972 \times 10^{-5}</td>
<td>0.073</td>
<td>1.23</td>
<td>0.00038</td>
<td>709821</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity to correlation $\rho$. We know that an accurate and reliable algorithm will yield a different probability estimate $\hat{\ell}$ for different values of $\rho$, and in particular, as the correlation coefficient $\rho$ increases and the dependence amongst $X_1, \ldots, X_d$ increases in strength, the corresponding probability $\ell(\rho)$ has to become larger and larger, reflecting, for example, the increased risk of bankruptcy of an insurer who holds a portfolio of highly interdependent (positively correlated) loans with little diversification.

With the last comment in mind, note the second column of Table 1, which shows the normalizing constant $\ell_1$ of the reference pdf. Although, the reference value has the property that $\ell_2 \downarrow \ell_1$ as $\gamma \uparrow \infty$, (Asmussen et al. 2011) [Proposition 1 and Theorem 1], the numerical experiments suggest that it does not capture well the effect of the correlation coefficient $\rho$ for finite $\gamma$. This is because the M-estimates (see
third column) are significantly larger than $\ell_1$ for large values of $\gamma$. Further, the ISVE estimates are not much different from the asymptotic approximation $\ell_1$. In fact, the lack of sensitivity of the ISVE estimator to $\rho$ deteriorates as $\gamma$ becomes larger as the next experiment illustrates.

The figure below shows the effect of increasing $\rho$ from $1 - 0.5$ to $1 - 0.5^{10}$ on the value of $\hat{\ell}$ for $\gamma = 5 \times 10^{15}$. The empty circles are the estimates obtained by the proposed method and the filled red dots lying on a line are the estimates obtained by the ISVE algorithm, both using the same simulation effort of $n = 10^6$. The proposed algorithm behaves as expected: a higher correlation $\rho$ increases the probability $\ell$. The ISVE algorithm on the other hand does not capture the effect of the correlation parameter very well. In fact, the ISVE estimator yields the same value as the asymptotic approximation $\ell_1$ (we observed the same phenomenon regarding the estimation of the tail of the maximum of correlated Gaussian densities (Botev et al. 2015)).

Figure 1: Effect of the correlation parameter $\rho = 1 - 0.5^c$, $c = 1, \ldots, 10$ on the rare-event probability $\ell(\rho)$. The empty circles represent the M-estimates and the dots lying on the line are the ISVE estimates.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we described a maximum-likelihood type estimator for the Monte Carlo estimation of certain rare-event probabilities. The M-estimator has the advantage over the standard importance sampling estimator in that the support and tail conditions on the proposal densities are much relaxed. Interestingly, the well-known fixed-effort splitting estimator can be viewed as an M-estimator.

We gave a numerical example of estimating the tail distribution of the sum of correlated log-normal random variables, in which the M-estimator is competitive. Space considerations permit us to consider only a limited number of numerical examples in this article, and a number of other successful applications of the method are documented elsewhere in (Huang and Botev 2013).

As future work, we plan to exploit the possibility of providing error estimates for the splitting estimator using maximum likelihood theory. A proper analysis of the error will have to take into account the error from using (approximate) MCMC sampling from the mixture density $\tilde{f}$. 367
Another issue, which requires further exploration, is the optimal choice of the mixture components \( \{\lambda_i\} \). So far, we have assumed that these are equal in our numerical experiments, but this choice is most certainly not optimal. Recent work on optimal design and MCMC sampling may provide clues as to the optimal choice of these weights (Doss and Tan 2014).

Finally, beyond numerical experiments, we have said nothing about the theoretical efficiency of the estimator \( \hat{\ell} \) with respect to the rarity parameter \( \gamma \). Future work must address the possibility of bounded relative error estimators in line with the results in (Gudmundsson and Hult 2014).

REFERENCES


**AUTHOR BIOGRAPHIES**

**ZDRAVKO I. BOTEV** is a Senior Lecturer at the School of Mathematics and Statistics at the University of New South Wales in Sydney, Australia. He obtained his Ph.D. in Mathematics from The University of Queensland, Australia, in 2010. His research interests include splitting and adaptive importance sampling methods for rare-event simulation. For more information, visit his webpage: [http://web.maths.unsw.edu.au/~zdravkobotev/](http://web.maths.unsw.edu.au/~zdravkobotev/)

**AD RIDDER** is an Associate Professor of operations research at the Department Econometrics and Operations Research, Vrije University, Amsterdam, The Netherlands. He obtained his Ph.D. in applied probability from Leiden University in 1987. His affiliations after his graduation include the University of California at Berkeley and the Rotterdam School of Management of Erasmus University. His main research interest is in efficient simulation methodologies for complex stochastic systems, with a focus on rare-event-related techniques. He is co-author of the book *Fast Sequential Monte Carlo Methods for Counting and Optimization*, Wiley, 2013.