Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

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 suggests that for this example the method is competitive.

1 INTRODUCTION

Suppose we wish to estimate a probability of the form

$$\ell = \mathbb{P}(S(\boldsymbol{X}) > \gamma), \quad \boldsymbol{X} = (X_1, \ldots, X_d),$$

where: (a) $S : \mathbb{R}^d \mapsto \mathbb{R}$ is the so-called importance function; (b) $X_1, \ldots, X_d$ are random variables with joint density $f(\boldsymbol{x})$; and (c) γ is a threshold, which may be large enough to make ℓ 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: $S(\boldsymbol{X}) = \exp(X_1) + \cdots + \exp(X_d)$, where $\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$, (Kortschak and Hashorva 2013, Asmussen et al. 2014, Laub et al. 2015).

Recently, a number of methods have been proposed for the estimation of ℓ 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 ℓ, 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 ℓ, 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 splitting 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 ℓ, it is convenient to think of ℓ as a normalization constant ℓ_s of the conditional density (here $\mathbb{I}\{\cdot\}$ is the indicator function of an event)

$$f_s(\boldsymbol{x}) = \frac{w_s(\boldsymbol{x})}{\ell_s} \stackrel{\text{def}}{=} \frac{f(\boldsymbol{x})\mathbb{I}\{S(\boldsymbol{x}) > \gamma\}}{\ell_s} .$$

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 $f_1(\boldsymbol{x}) = w_1(\boldsymbol{x})/\ell_1$ be another *reference* or *importance sampling* density whose normalizing constant ℓ_1 is known. Then, the standard importance sampling estimator with n samples is

$$\ell_s^* = \frac{1}{n}\sum_{j=1}^{n} \frac{w_s(\boldsymbol{X}_j)}{f_1(\boldsymbol{X}_j)}, \qquad \boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim f_1 . \tag{1}$$

For acceptable performance, see (Kroese et al. 2011), we not only need the condition $\{\boldsymbol{x} : f_1(\boldsymbol{x}) > 0\} \supseteq \{\boldsymbol{x} : f_s(\boldsymbol{x}) > 0\}$, 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 $\bar{f} \equiv \lambda_1 f_1 + \lambda_s f_s$, where $\lambda_1 + \lambda_s = 1$ are some weight fixed in advance. A plausible importance-sampling-type estimator then looks like:

$$\hat{\ell}_s = \frac{1}{n}\sum_{j=1}^{n} \frac{w_s(\boldsymbol{X}_j)}{\lambda_1 f_1(\boldsymbol{X}_j) + \lambda_s \underbrace{w_s(\boldsymbol{X}_j)/\hat{\ell}_s}_{\approx f_s}}, \qquad \boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim \bar{f} \tag{2}$$

where the unknown normalizing constant ℓ_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 ℓ_s^*? Compared with the traditional importance sampling estimator, ℓ_s^*, which requires the quite restrictive conditions on the tail and support of f_1 ($\{\boldsymbol{x} : f_1(\boldsymbol{x}) > 0\} \supseteq \{\boldsymbol{x} : f_s(\boldsymbol{x}) > 0\}$ and $\mathbb{E}_{f_s} f_s(\boldsymbol{X})/f_1(\boldsymbol{X}) < \infty$), the tail and support restrictions on f_1 in the estimator $\hat{\ell}_s$ are relaxed to the much weaker $\{\boldsymbol{x} : f_1(\boldsymbol{x}) \times f_s(\boldsymbol{x}) > 0\} \neq \emptyset$. 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 $f_1(\boldsymbol{x}) \propto f(\boldsymbol{x})\mathbb{I}\{\max_i \exp(X_i) > \gamma\}$, then this f_1 cannot be used in the standard importance sampling estimator ℓ_s^*, but it can be used in the estimator $\hat{\ell}_s$, because the sets $\{\boldsymbol{x} : \exp(x_1) + \cdots + \exp(x_d) > \gamma\}$ and $\{\boldsymbol{x} : \max_i \exp(x_i) > \gamma\}$ 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)

$$f_t(\boldsymbol{x}) = \frac{w_t(\boldsymbol{x})}{\ell_t}, \quad t = 1, \dots, s ,$$

where ℓ_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

$$\mathbb{P}(w_i(\boldsymbol{X}) \times w_j(\boldsymbol{X}) > 0) > 0, \tag{3}$$

then the connectivity condition says that there exists a path between any two nodes of the graph. Recall that we are interested in estimating ℓ_s and we assume that ℓ_1 is known (we call f_1 a reference density), and without loss of generality, $\ell_1 = 1$. To estimate ℓ_s, we may simulate from each density as follows

$$\boldsymbol{X}_{t,1}, \ldots, \boldsymbol{X}_{t,n_t} \sim f_t(\boldsymbol{x}), \quad t = 1, \ldots, s\ ,$$

and collect the pooled sample as $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, 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

$$\bar{f}(\boldsymbol{x}) = \frac{1}{n}\sum_{t=1}^{s} n_t f_t(\boldsymbol{x}) = \sum_{t=1}^{s} \lambda_t f_t(\boldsymbol{x}), \quad \lambda_t \stackrel{\text{def}}{=} n_t/n\ .$$

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

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

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

$$\hat{\boldsymbol{z}} = \operatorname*{argmax}_{\boldsymbol{z}} \mathcal{D}_n(\boldsymbol{z}), \tag{4}$$

where we have the likelihood-lookalike objective function

$$\mathcal{D}_n(\boldsymbol{z}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} \log\left(\sum_{k=1}^{s} w_k(\boldsymbol{X}_j)\ \exp(z_k)\right) - \sum_{k=1}^{s} \lambda_k z_k \tag{5}$$

and $\mathcal{D}(\boldsymbol{z}) \stackrel{\text{def}}{=} \int \bar{f}(\boldsymbol{x}) \log\left(\sum_{k=1}^{s} w_k(\boldsymbol{x})\ \exp(z_k)\right) \mathrm{d}\boldsymbol{x} - \sum_{k=1}^{s} \lambda_k z_k$.

It may not be clear why the solution of this program yields a sensible estimator of the true $\boldsymbol{z}$ or $\boldsymbol{\ell}$. The first reason why this estimator makes sense is the following consistency result. (Here $X_n \stackrel{\mathbb{P}}{\to} X$ means that for any ϵ, δ 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 $\bar{f}$ (usually accomplished via MCMC) is such that a weak law of large numbers applies, $\mathcal{D}_n(\boldsymbol{z}) \stackrel{\mathbb{P}}{\to} \mathcal{D}(\boldsymbol{z})$, uniformly in $\boldsymbol{z}$, then $\hat{\boldsymbol{z}} \stackrel{\mathbb{P}}{\to} \boldsymbol{z}$ 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, $\mathcal{D}_n(\boldsymbol{z}) \stackrel{\mathbb{P}}{\to} \mathcal{D}(\boldsymbol{z})$, that for any $\tilde{\boldsymbol{z}} \neq \boldsymbol{z}$ (the first component is also $\tilde{z}_1 = \hat{z}_1 = z_1$):

$$\begin{aligned}\mathcal{D}_n(\tilde{\boldsymbol{z}}) - \mathcal{D}_n(\boldsymbol{z}) &= \overbrace{\mathcal{D}_n(\tilde{\boldsymbol{z}}) - \mathcal{D}(\tilde{\boldsymbol{z}})}^{\stackrel{\mathbb{P}}{\to} 0} + \overbrace{\mathcal{D}(\boldsymbol{z}) - \mathcal{D}_n(\boldsymbol{z})}^{\stackrel{\mathbb{P}}{\to} 0} + \mathcal{D}(\tilde{\boldsymbol{z}}) - \mathcal{D}(\boldsymbol{z}) \\ &\stackrel{\mathbb{P}}{\to} 0 + 0 + \underbrace{\mathcal{D}(\tilde{\boldsymbol{z}}) - \mathcal{D}(\boldsymbol{z})}_{\alpha(\tilde{\boldsymbol{z}}, \boldsymbol{z})},\end{aligned}$$

where we have denoted the last expression by

$$\alpha(\tilde{\boldsymbol{z}}, \boldsymbol{z}) \stackrel{\text{def}}{=} \int \bar{f}(\boldsymbol{x}) \log \left(\frac{\sum_{k=1}^{s} w_k(\boldsymbol{x}) \, \exp(\tilde{z}_k)}{\sum_{k=1}^{s} w_k(\boldsymbol{x}) \, \exp(z_k)} \right) \mathrm{d}\boldsymbol{x} - \sum_{k=1}^{s} \lambda_k (\tilde{z}_k - z_k)$$

Since $\bar{f}(\boldsymbol{x}) = \sum_{t=1}^{s} \lambda_t f_t(\boldsymbol{x}) = \sum_{t=1}^{s} \exp(z_t) w_t(\boldsymbol{x})$, we can apply Jensen's inequality:

$$\begin{aligned}
\int \bar{f}(\boldsymbol{x}) \log \left(\frac{\sum_{k=1}^{s} w_k(\boldsymbol{x}) \, \exp(\tilde{z}_k)}{\sum_{k=1}^{s} w_k(\boldsymbol{x}) \, \exp(z_k)} \right) \mathrm{d}\boldsymbol{x} &= \int \bar{f}(\boldsymbol{x}) \log \left(\frac{\sum_{k=1}^{s} w_k(\boldsymbol{x}) \, \exp(\tilde{z}_k)}{\bar{f}(\boldsymbol{x})} \right) \mathrm{d}\boldsymbol{x} \\
\text{Jensen's inequality} \qquad &\leq \log \left(\int \sum_{k=1}^{s} w_k(\boldsymbol{x}) \, \exp(\tilde{z}_k) \mathrm{d}\boldsymbol{x} \right) \\
&\leq \log \left(\sum_{k=1}^{s} \exp(\tilde{z}_k) \ell_k \right) = \log \left(\sum_{k=1}^{s} \lambda_k \exp(\tilde{z}_k - z_k) \right)
\end{aligned}$$

Therefore,

$$\mathcal{D}_n(\tilde{\boldsymbol{z}}) - \mathcal{D}_n(\boldsymbol{z}) \stackrel{\mathbb{P}}{\to} \alpha(\tilde{\boldsymbol{z}}, \boldsymbol{z}) \leq \log \left(\sum_{k=1}^{s} \lambda_k \exp(\tilde{z}_k - z_k) \right) - \sum_{k=1}^{s} \lambda_k (\tilde{z}_k - z_k).$$

Next, another application of Jensen's inequality with the distribution $\{\lambda_k\}$ yields

$$\beta(\tilde{\boldsymbol{z}}, \boldsymbol{z}) \stackrel{\text{def}}{=} \log \left(\sum_{k=1}^{s} \lambda_k \exp(\tilde{z}_k - z_k) \right) - \sum_{k=1}^{s} \lambda_k (\tilde{z}_k - z_k) \geq 0.$$

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 $\tilde{\boldsymbol{z}} = \boldsymbol{z} + c$, where c is an arbitrary constant. In our case $c = 0$, because $\tilde{z}_1 = z_1$ by construction. In other words, $\beta(\tilde{\boldsymbol{z}}, \boldsymbol{z}) > 0$ for $\boldsymbol{z} \neq \tilde{\boldsymbol{z}}$.

It follows that if $\boldsymbol{z} \neq \tilde{\boldsymbol{z}}$, then with increasing probability as $n \uparrow \infty$, $\mathcal{D}_n(\tilde{\boldsymbol{z}}) - \mathcal{D}_n(\boldsymbol{z})$ will be upper bounded by a strictly positive constant β. That is, as $n \uparrow \infty$ and $\boldsymbol{z} \neq \tilde{\boldsymbol{z}}$ we have $\mathbb{P}(\mathcal{D}_n(\boldsymbol{z}) > \mathcal{D}_n(\tilde{\boldsymbol{z}})) \uparrow 1$.

Hence, using the fact that $\mathcal{D}_n(\tilde{\boldsymbol{z}}) \leq \mathcal{D}_n(\hat{\boldsymbol{z}})$ for all $\tilde{\boldsymbol{z}}$, because $\hat{\boldsymbol{z}}$ is, by construction, a global maximizer of the (almost surely) concave $\mathcal{D}_n$, we have as $n \uparrow \infty$ and any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}(\|\hat{\boldsymbol{z}} - \boldsymbol{z}\| > \epsilon) &= \mathbb{P}(\|\hat{\boldsymbol{z}} - \boldsymbol{z}\| > \epsilon, \mathcal{D}_n(\tilde{\boldsymbol{z}}) \leq \mathcal{D}_n(\hat{\boldsymbol{z}}), \text{ for all } \tilde{\boldsymbol{z}}) \\
&\leq \mathbb{P}(\|\hat{\boldsymbol{z}} - \boldsymbol{z}\| > \epsilon, \mathcal{D}_n(\boldsymbol{z}) \leq \mathcal{D}_n(\hat{\boldsymbol{z}})) \\
&\leq \mathbb{P}(\hat{\boldsymbol{z}} \neq \boldsymbol{z}, \mathcal{D}_n(\boldsymbol{z}) \leq \mathcal{D}_n(\hat{\boldsymbol{z}})) \\
&\leq 1 - \mathbb{P}(\mathcal{D}_n(\boldsymbol{z}) > \mathcal{D}_n(\tilde{\boldsymbol{z}}), \tilde{\boldsymbol{z}} \neq \boldsymbol{z}) \to 0
\end{aligned}$$

□

A second way to see that estimator (4) is sensible is to solve for the gradient of $\mathcal{D}_n$ being zero: $\nabla \mathcal{D}_n(\hat{\boldsymbol{z}}) = \mathbf{0}$. Rewriting this nonlinear system explicitly in $\boldsymbol{\ell}$ gives the system of *moment-matching* equations:

$$\hat{\ell}_t = \frac{1}{n} \sum_{j=1}^{n} \frac{w_t(\boldsymbol{X}_j)}{\sum_{k=1}^{s} w_k(\boldsymbol{X}_j) \lambda_k / \hat{\ell}_k}, \quad t = 2, \ldots, s, \tag{6}$$

which appear to be the sample versions of the identities:

$$\ell_t = \mathbb{E}_{\bar{f}} \left[\frac{w_t(\boldsymbol{X})}{\sum_{k=1}^{s} w_k(\boldsymbol{X}) \lambda_k / \ell_k} \right] .$$

The moment-matching equations (6) also suggest an iterative method for solving the nonlinear system $\nabla \mathcal{D}_n(\hat{\boldsymbol{z}}) = \mathbf{0}$, namely, via the following iterative procedure.

Algorithm 1 : Jacobi fixed-point iteration

Require: Initial $\hat{\boldsymbol{\ell}} = (1, \ell_2, \ldots, \ell_s)^\top$
  Set $\epsilon = \infty$ and $\boldsymbol{\ell} \leftarrow \hat{\boldsymbol{\ell}}$
  while $\epsilon > 10^{-5}$ **do**
    for $i = 2, \ldots, s$ **do**
      $\ell_i \leftarrow \frac{1}{n} \sum_{j=1}^{n} \frac{w_i(\boldsymbol{X}_j)}{\sum_{k=1}^{s} w_k(\boldsymbol{X}_j)\, \lambda_k/\ell_k}$
    $\epsilon \leftarrow \max_i \frac{|\ell_i - \hat{\ell}_i|}{\ell_i}$
  return The vector of estimated probabilities $\hat{\boldsymbol{\ell}} \leftarrow \boldsymbol{\ell}$.

In the case of iid sampling from $\bar{f}$, we can deduce from the multivariate delta method the maximum-likelihood-type asymptotics $\sqrt{n}(\hat{\boldsymbol{z}} - \boldsymbol{z}) \xrightarrow{\mathrm{d}} \mathsf{N}(\mathbf{0}, \mathrm{F}^{-1})$, where $\mathrm{F} = \mathbb{E}_{\bar{f}} \nabla^2 \mathcal{D}_n$ plays the role of a "Fisher" information matrix. Since the $\boldsymbol{\ell}$ equals $\boldsymbol{z}$ on a logarithmic scale, this suggests the estimator, $\hat{\mathrm{F}}/n$, of the relative error of $\hat{\boldsymbol{\ell}}$. In practice, we have to use MCMC for simulation from $\bar{f}$, and for this reason, we estimate the relative error using the batch means method (Kroese et al. 2011, Algorithm 8.4) without any burn-in. We will give an example in the numerical section.

In the next section we consider the relationship between the fixed-effort splitting (Botev and Kroese 2012, Botev and Kroese 2008, Botev 2009) and the M-estimator, and in particular show that the splitting estimator is an analytical solution of the moment-matching equations (6).

4 SPLITTING AND M-ESTIMATOR

Consider the special case in which all the densities $f_t(\boldsymbol{x})$ are of the form:

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

where $-\infty = \gamma_1 < \gamma_2 < \cdots < \gamma_s = \gamma$ (note that f_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 $\mathbb{P}(S(\boldsymbol{X}) > \gamma_j, S(\boldsymbol{X}) > \gamma_i) = \mathbb{P}(S(\boldsymbol{X}) > \max\{\gamma_i, \gamma_j\}) \geq \ell > 0$. 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 $\mathbb{I}_{k,j} = \mathbb{I}\{S(\boldsymbol{X}_j) > \gamma_k\}$. Then, the unique solution of (6) is:

$$\hat{\ell}_t = \ell_1 \prod_{k=2}^{t} \frac{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k,j}}{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k-1,j}}, \quad t = 2, \ldots, s\,. \tag{7}$$

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 $t = 2, \dots, s-1$. Expanding the size of the matrix to $s \times (n_1 + \cdots + n_s)$ gives that (6) is identical to

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

from where we have $n_s + \ell_s \sum_{k \neq s} \frac{n_k}{\ell_k} = \sum_{j=1}^{n} \mathbb{I}_{s,j} = n_s + \sum_{j=1}^{n_1+\cdots+n_{s-1}} \mathbb{I}_{s,j}$ and hence the only solution to (6) is:

$$\ell_s = \frac{\sum_{j=1}^{n_1+\cdots+n_{s-1}} \mathbb{I}_{s,j}}{\sum_{k \neq s} \frac{n_k}{\ell_k}} = \ell_1 \prod_{k=2}^{s} \frac{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k,j}}{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k-1,j}},$$

because a direct calculation from left to right shows that

$$\begin{aligned}
\frac{n_1}{\ell_1} + \frac{n_2}{\ell_2} + \cdots + \frac{n_{s-1}}{\ell_{s-1}} &= \frac{n_1}{\ell_1} \frac{\sum_{j=1}^{n_1+n_2} \mathbb{I}_{2,j}}{\sum_{j=1}^{n_1} \mathbb{I}_{2,j}} + \frac{n_3}{\ell_3} + \cdots + \frac{n_{s-1}}{\ell_{s-1}} \\
&= \frac{n_1}{\ell_1} \frac{\sum_{j=1}^{n_1+n_2} \mathbb{I}_{2,j}}{\sum_{j=1}^{n_1} \mathbb{I}_{2,j}} \frac{\sum_{j=1}^{n_1+n_2+n_3} \mathbb{I}_{3,j}}{\sum_{j=1}^{n_1+n_2} \mathbb{I}_{3,j}} + \frac{n_4}{\ell_4} + \cdots + \frac{n_{s-1}}{\ell_{s-1}} \\
&= \frac{n_1}{\ell_1} \prod_{k=2}^{s-1} \frac{\sum_{j=1}^{n_1+\cdots+n_k} \mathbb{I}_{k,j}}{\sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k,j}} = \frac{n_1}{\ell_1} \frac{\prod_{k=2}^{s-1} \sum_{j=1}^{n_1+\cdots+n_k} \mathbb{I}_{k,j}}{\prod_{k=2}^{s-1} \sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k,j}} \\
&= \frac{n_1}{\ell_1} \frac{\prod_{k=3}^{s} \sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k-1,j}}{\prod_{k=2}^{s-1} \sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k,j}} = \frac{1}{\ell_1} \frac{\prod_{k=2}^{s} \sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k-1,j}}{\prod_{k=2}^{s-1} \sum_{j=1}^{n_1+\cdots+n_{k-1}} \mathbb{I}_{k,j}}
\end{aligned}$$

□

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 ($s = 2$)

$$f_2(\boldsymbol{x}) = \frac{f(\boldsymbol{x})\, \mathbb{I}\{S(\boldsymbol{x}) \geq \gamma\}}{\ell_2},$$

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

We now estimate ℓ_2 via the estimator (2) with $s = 2$ and reference density

$$f_1(\boldsymbol{x}) = \frac{f(\boldsymbol{x}) \sum_{j=1}^{d} \mathbb{I}\{\exp(x_i) > \gamma\}}{\ell_1}, \qquad \ell_1 \stackrel{\text{def}}{=} \sum_{j=1}^{d} \mathbb{P}\left(\exp(X_i) > \gamma\right), \qquad (X_1, \dots, X_d) \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)\,.$$

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

$$f_1(\boldsymbol{x}) = \sum_{j=1}^{d} \frac{\mathbb{P}(X_j > \log\gamma)}{\ell_1} \frac{f(\boldsymbol{x})\mathbb{I}\{x_j > \log\gamma\}}{\mathbb{P}(X_j > \log\gamma)}.$$

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 $f(\boldsymbol{x} \,|\, x_j)$. 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

$$f(x_i \,|\, \boldsymbol{x}_{-i}) \times \mathbb{I}\{x_i > \ln(\gamma - \textstyle\sum_{j\neq i} \exp(x_j))\}$$

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 $\{\boldsymbol{x} : S(\boldsymbol{x}) > \gamma, \max_i \exp(x_i) < \gamma\}$. 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 $S(\boldsymbol{X}) > \gamma$ always under this simulation scheme):

$$\begin{aligned}\hat{\ell}_2 &= \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \dfrac{w_2(\boldsymbol{X}_j)}{\lambda_1\, w_1(\boldsymbol{X}_j)/\ell_1 + \lambda_2\, w_2(\boldsymbol{X}_j)/\hat{\ell}_2} = \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \dfrac{1}{\lambda_1 \frac{w_1(\boldsymbol{X}_j)}{f(\boldsymbol{X}_j)}/\ell_1 + \lambda_2/\hat{\ell}_2} \\ &= \frac{p_0}{\frac{0\,n_1}{\ell_1} + \frac{n_2}{\hat{\ell}_2}} + \frac{p_1}{\frac{n_1}{\ell_1} + \frac{n_2}{\hat{\ell}_2}} + \frac{p_2}{\frac{2\,n_1}{\ell_1} + \frac{n_2}{\hat{\ell}_2}} + \cdots + \frac{p_d}{\frac{d\,n_1}{\ell_1} + \frac{n_2}{\hat{\ell}_2}},\end{aligned}$$

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

$$p_0 + \frac{p_1}{\hat{\ell}_2 \frac{n_1}{n_2\,\ell_1} + 1} + \frac{p_2}{\hat{\ell}_2 \frac{2\,n_1}{n_2\,\ell_1} + 1} + \cdots + \frac{p_d}{\hat{\ell}_2 \frac{d\,n_1}{n_2\,\ell_1} + 1} = n_2\ .$$

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 `fzero.m` in Matlab to solve the equation without difficulty.

Comparison between M-est. and ISVE. Tables 1 and 2 show the estimates of ℓ for various values of γ and ρ. 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 $\hat{\boldsymbol{\ell}}$ (which is asymptotically the variance of $\hat{\boldsymbol{z}}$), 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 (WNRV), which takes into account the computational time for each method. The WNRV is defined as $\tau \times \mathrm{Var}(\hat{\ell})/\ell^2$, where τ is the CPU time.

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 $n = n_1 + n_2 = 5\times 10^5$.

				relative error %		WNRV	
γ	ℓ_1	M-est.	ISVE estim.	M-est.	ISVE	M-est.	ISVE
5×10^4	0.000355	0.000409	0.000406	0.23	1.71	0.00044	15248
5×10^5	1.794×10^{-5}	2.212×10^{-5}	2.177×10^{-5}	0.23	3.09	0.00043	50267
5×10^6	5.586×10^{-7}	7.156×10^{-7}	6.807×10^{-7}	0.23	5.32	0.00042	1.4×10^5
5×10^7	1.057×10^{-8}	1.384×10^{-8}	1.444×10^{-8}	0.23	11.74	0.00042	7.2×10^5
5×10^8	1.205×10^{-10}	1.590×10^{-10}	1.254×10^{-10}	0.23	2.35	0.00042	29064
5×10^9	8.230×10^{-13}	1.086×10^{-12}	3.781×10^{-12}	0.23	76.90	0.00040	3.13×10^7
5×10^{10}	3.347×10^{-15}	4.372×10^{-15}	3.346×10^{-15}	0.22	0.10	0.00040	56.12
5×10^{11}	8.087×10^{-18}	1.046×10^{-17}	8.083×10^{-18}	0.22	0.024	0.00039	2.99
5×10^{12}	1.158×10^{-20}	1.483×10^{-20}	1.158×10^{-20}	0.22	0.0018	0.00039	0.016
5×10^{13}	9.827×10^{-24}	1.245×10^{-23}	1.641×10^{-23}	0.22	40.12	0.00039	8.38×10^6
5×10^{14}	4.930×10^{-27}	6.170×10^{-27}	5.028×10^{-27}	0.22	1.94	0.00039	19790
5×10^{15}	1.462×10^{-30}	1.804×10^{-30}	1.462×10^{-30}	0.22	0.00037	0.00038	0.00073
5×10^{16}	2.562×10^{-34}	3.123×10^{-34}	2.563×10^{-34}	0.22	0.00020	0.00038	0.00020
5×10^{17}	2.651×10^{-38}	3.198×10^{-38}	2.652×10^{-38}	0.22	0.00010	0.00037	5.21×10^{-5}

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

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

Sensitivity to correlation ρ. We know that an accurate and reliable algorithm will yield a different probability estimate $\hat{\ell}$ for different values of ρ, and in particular, as the correlation coefficient ρ 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 ℓ_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 ρ for finite γ. This is because the M-estimates (see

third column) are significantly larger than ℓ_1 for large values of γ. Further, the ISVE estimates are not much different from the asymptotic approximation ℓ_1. In fact, the lack of sensitivity of the ISVE estimator to ρ deteriorates as γ becomes larger as the next experiment illustrates.

The figure below shows the effect of increasing ρ 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 ρ increases the probability ℓ. 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 ℓ_1 (we observed the same phenomenon regarding the estimation of the tail of the maximum of correlated Gaussian densities (Botev et al. 2015)).

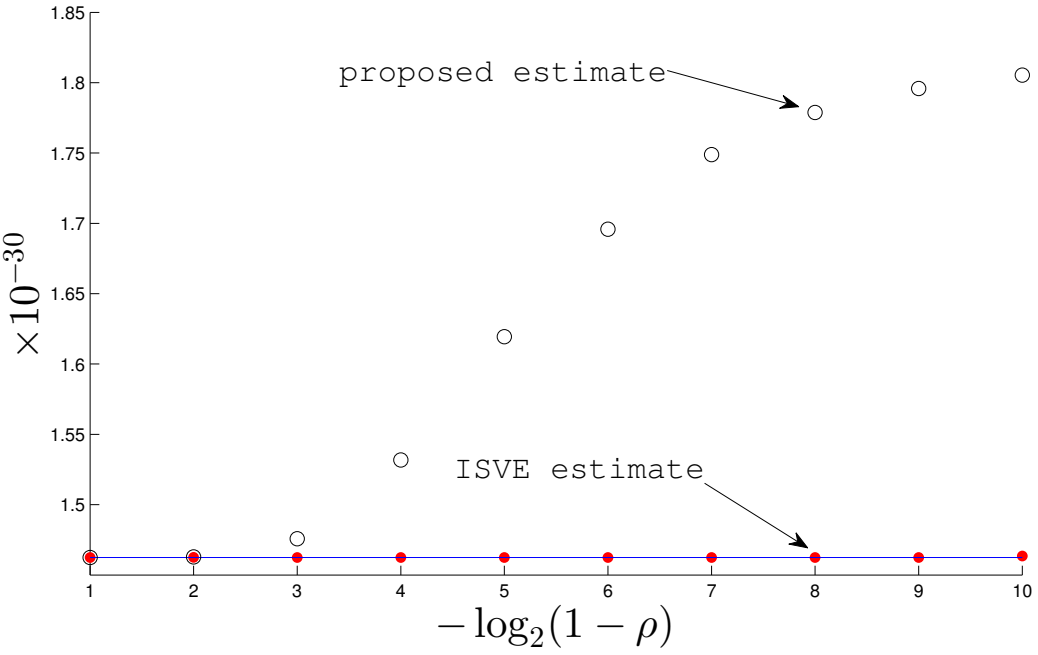

Figure 1: Effect of the correlation parameter $\rho = 1 - 0.5^c,\ c = 1, \ldots, 10$ on the rare-event probability $\ell(\rho)$. The empty circles represent the M-estimates and the dots lying on the line are the ISVE estimates.

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 $\bar{f}$.

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 γ. Future work must address the possibility of bounded relative error estimators in line with the results in (Gudmundsson and Hult 2014).

REFERENCES

Asmussen, S., J. Blanchet, S. Juneja, and L. Rojas-Nandayapa. 2011. "Efficient Simulation of Tail Probabilities of Sums of Correlated Lognormals". *Annals of Operations Research* 189 (1): 5–23.

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York: Springer-Verlag.

Asmussen, S., J. L. Jensen, and L. Rojas-Nandayapa. 2014. "Exponential Family Techniques for the Lognormal Left Tail". *arXiv preprint arXiv:1403.4689*.

Botev, Z. I. 2009. "Splitting Methods for Efficient Combinatorial Counting and Rare-Event Probability Estimation". PhD thesis: The University of Queensland Library.

Botev, Z. I., and D. P. Kroese. 2008. "An Efficient Algorithm for Rare-event Probability Estimation, Combinatorial Optimization, and Counting". *Methodology and Computing in Applied Probability* 10 (4): 471–505.

Botev, Z. I., and D. P. Kroese. 2012. "Efficient Monte Carlo Simulation via the Generalized Splitting method". *Statistics and Computing* 22 (1): 1–16.

Botev, Z. I., P. L'Ecuyer, and B. Tuffin. 2011. "An Importance Sampling Method Based on a One-step Look-ahead Density from a Markov Chain". In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasley, J. Himmelspach, K. P. White, and M. Fu, 528–539. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Botev, Z. I., P. L'Ecuyer, and B. Tuffin. 2013. "Markov Chain Importance Sampling with Applications to Rare-event Probability Estimation". *Statistics and Computing* 23 (2): 271–285.

Botev, Z. I., M. Mandjes, and A. Ridder. 2015. "Tail Distribution of the Maximum of Correlated Gaussian Random Variables". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 633–642. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Botev, Z. I., A. Ridder, and L. Rojas-Nandayapa. 2016. "Semiparametric Cross Entropy For Rare-Event Simulation". *Journal of Applied Probability* 53 (3).

Doss, H., and A. Tan. 2014. "Estimates and Standard Errors for Ratios of Normalizing Constants from Multiple Markov Chains via Regeneration". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (4): 683–712.

Gill, R. D., Y. Vardi, and J. A. Wellner. 1988. "Large Sample Theory of Empirical Distributions in Biased Sampling Models". *The Annals of Statistics* 16 (3): 1069–1112.

Gudmundsson, T., and H. Hult. 2014. "Markov Chain Monte Carlo for Computing Rare-event Probabilities for a Heavy-tailed Random Walk". *Journal of Applied Probability* 51 (2): 359–376.

Huang, A., and Z. I. Botev. 2013. "Rare-event Probability Estimation via Empirical Likelihood Maximization". *arXiv preprint arXiv:1312.3027*.

Kong, A., P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan. 2003. "A Theory of Statistical Models for Monte Carlo Integration". *Journal of the Royal Statistical Society, Series B* 65 (3): 585–618.

Kortschak, D., and E. Hashorva. 2013. "Efficient Simulation of Tail Probabilities for Sums of Log-elliptical Risks". *Journal of Computational and Applied Mathematics* 247:53–67.

Kroese, D. P., T. Taimre, and Z. I. Botev. 2011. *Handbook of Monte Carlo Methods*, Volume 706. Wiley.

Laub, P. J., S. Asmussen, J. L. Jensen, and L. Rojas-Nandayapa. 2015. "Approximating the Laplace Transform of the Sum of Dependent Lognormals". *arXiv preprint arXiv:1507.03750*.
Vardi, Y. 1985. "Empirical Distributions in Selection Bias Models". *The Annals of Statistics* 13 (1): 178–203.

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/

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.