Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

EFFICIENT PROBABILITY ESTIMATION AND SIMULATION OF THE TRUNCATED MULTIVARIATE STUDENT-t DISTRIBUTION

Zdravko I. Botev

School of Mathematics and Statistics
The University of New South Wales
Sydney, NSW 2052, AUSTRALIA

Pierre L'Ecuyer

DIRO, Université de Montreal
C.P. 6128, Succ. Centre-Ville
Montréal (Québec), H3C 3J7, CANADA

ABSTRACT

We propose an exponential tilting method for the accurate estimation of the probability that a random vector with multivariate student-t distribution falls in a convex polytope. The method can also be used to simulate exactly from the corresponding truncated multivariate student-t distribution, thus providing an alternative to approximate Markov Chain Monte Carlo simulation. Numerical experiments show that the suggested method is significantly more accurate and reliable than its competitors.

1 INTRODUCTION

Let $\mathbf{X} \in \mathbb{R}^d$ be distributed according to the multivariate student-t distribution with $\nu > 0$ degrees of freedom. The density of $\mathbf{X}$ is given by

$$c_1 \times \left(1 + \frac{1}{\nu}\|\mathbf{x}\|^2\right)^{-(\nu+d)/2}$$

over $\mathbb{R}^d$, where

$$c_1 = \frac{\Gamma((d+\nu)/2)}{(\pi\nu)^{d/2}\Gamma(\nu/2)}$$

is a normalizing constant. We write $\mathbf{X} \sim \mathbf{t}_\nu$. We are interested in two closely related problems. The first one is estimating the probability

$$\ell = \mathbb{P}(\mathbf{l} \leqslant C\mathbf{X} \leqslant \mathbf{u}) = \int_{\mathbb{R}^d} c_1 \left(1 + \frac{1}{\nu}\|\mathbf{x}\|^2\right)^{-(\nu+d)/2} \mathbb{I}\{\mathbf{l} \leqslant C\mathbf{x} \leqslant \mathbf{u}\}\mathrm{d}\mathbf{x}, \tag{1}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function and C is a $d \times d$ full rank matrix. The second problem is to simulate exactly from the truncated (or conditional) density:

$$f(\mathbf{x}) = \frac{c_1 \times \left(1 + \frac{1}{\nu}\|\mathbf{x}\|^2\right)^{-(\nu+d)/2} \times \mathbb{I}\{\mathbf{l} \leqslant C\mathbf{x} \leqslant \mathbf{u}\}}{\ell}. \tag{2}$$

Both of these problems arise frequently in statistical applications; see Genz and Bretz (2002), Genz (2004), Genz and Bretz (2009), and the references therein.

The purpose of this article is to propose a method for estimating (1) that is more reliable and efficient than the current state-of-the-art method of Genz (2004), also described in Genz and Bretz (2009), which is currently the default algorithm in MATLAB® and **R**. As a byproduct of the design of our algorithm, we can also sample from the conditional density (2) in high dimensions using an efficient acceptance-rejection scheme (Kroese et al. 2011, Chapter 3). Currently, the only practical method for simulation from the conditional (2), when ℓ is a rare-event probability, is by (approximate) Markov Chain Monte Carlo sampling;

see Yu and Tian (2011) and the references therein. Naturally, if ℓ is not a rare-event probability, say larger than 10^{-4}, then one can simulate exactly from (2) by simulating $\mathbf{X} \sim \mathbf{t}_\nu$ until the condition $\mathbf{l} \leqslant \mathbf{X} \leqslant \mathbf{u}$ is satisfied.

The idea of our method is to apply a suitable exponential tilting to the estimator proposed by Genz (2004). Exponential tilting is a popular way to construct a sampling density when applying importance sampling to estimate tail probabilities for light-tailed distributions (Bucklew 2004, Asmussen and Glynn 2007, L'Ecuyer et al. 2010). However, in our numerical simulations we observed significant efficiency gains even when they do not involve a tail probability setting, suggesting that exponential tilting is useful beyond its typical range of applications in large deviations.

We choose the tilting parameter by solving a convex optimization problem. This idea is similar to the recently proposed minimax exponential tilting for the multivariate normal distribution (Botev 2014), which relies on constructing a certain log-convex likelihood ratio. The main contribution of this article is to adapt the method for the multivariate normal to the multivariate student-*t* case using the fact that one can simulate a multivariate student-*t* vector by multiplying a multivariate normal vector with a suitable random scale variable. The adaptation is not a straightforward task, because we have to change the measure of the scale variable and most of the simple and obvious changes of measure cause a loss of the crucial log-convexity property of the likelihood ratio. Fortunately, we were able to find a change of measure of the scale variable that preserves this desirable log-convexity property. Another contribution in this article is the derivation of a simple nontrivial lower bound to ℓ.

The rest of the paper is organized as follows. We first describe in Section 2 the estimator originally proposed by Genz (2004). In Section 3 we describe our choice of exponential tilting. The exponential tilting approach allows us to both estimate ℓ accurately and simulate from the conditional density (2) in up to at least one hundred dimensions, with minimal additional computational overhead. Finally, in Section 4, we present numerical results demonstrating the superior practical performance of the new algorithms compared to the existing state-of-the-art.

2 AN ESTIMATOR OF ℓ BY SEPARATION OF VARIABLES

First, note that the multivariate student-*t* distribution forms a location scale family. In other words, if $\mathbf{X} \sim \mathbf{t}_\nu$ and $\mathbf{Y} = \breve{\mu} + A\mathbf{X}$, then we can write $\mathbf{Y} \sim \mathbf{t}_\nu(\breve{\mu}, AA^\top)$, where $\breve{\mu}$ and $AA^\top$ are the location and scale parameters, respectively. We thus have

$$\mathbb{P}(\breve{\mathbf{l}} \leqslant C\mathbf{Y} \leqslant \breve{\mathbf{u}}) = \mathbb{P}(\mathbf{l} \leqslant \breve{L}\mathbf{X} \leqslant \mathbf{u}),$$

where $\mathbf{X} \sim \mathbf{t}_\nu(\mathbf{0}, I_d) \equiv \mathbf{t}_\nu$, I_d is the d-dimensional identity matrix, $\mathbf{l} = \breve{\mathbf{l}} - C\breve{\mu}, \mathbf{u} = \breve{\mathbf{u}} - C\breve{\mu}$, and the matrix $\breve{L}$ satisfies $\breve{L}\breve{L}^\top = CAA^\top C^\top$. Hence, without loss of generality we need only consider the standardized versions (1) and (2).

Let L be the lower triangular Cholesky factor of the positive definite matrix $CC^\top$. Then, we can decompose (1) as follows:

$$\ell = \mathbb{P}(R\mathbf{l} \leqslant \sqrt{\nu} L\mathbf{Z} \leqslant R\mathbf{u}),$$

where R follows the χ_ν distribution with density

$$f_\nu(r) = \frac{\mathrm{e}^{-r^2/2 + (\nu-1)\ln r}}{2^{\nu/2-1}\Gamma(\nu/2)}, \quad \text{for } r > 0,$$

and $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, I_d)$ is a d-dimensional standard normal, independent of R. This is simply the well-known distributional result that $\sqrt{\nu}\mathbf{Z}/R \sim \mathbf{t}_\nu$; see Kroese et al. (2011), Chapter 3. Due to the lower triangular structure of L, the region $\mathscr{R} = \{(r,\mathbf{z}) : r\mathbf{l} \leqslant \sqrt{\nu}L\mathbf{z} \leqslant r\mathbf{u}\}$ can be decomposed into

$$\begin{aligned}
\tilde{l}_1(r) \stackrel{\text{def}}{=} \frac{r\,l_1\nu^{-1/2}}{L_{11}} &\leqslant \quad z_1 \quad \leqslant \frac{r\,u_1\nu^{-1/2}}{L_{11}} \stackrel{\text{def}}{=} \tilde{u}_1(r) \\
\tilde{l}_2(r,z_1) \stackrel{\text{def}}{=} \frac{r\,l_2\nu^{-1/2} - L_{21}z_1}{L_{22}} &\leqslant \quad z_2 \quad \leqslant \frac{r\,u_2\nu^{-1/2} - L_{21}z_1}{L_{22}} \stackrel{\text{def}}{=} \tilde{u}_2(r,z_1) \\
&\vdots \\
\tilde{l}_d(r,z_1,\ldots,z_{d-1}) \stackrel{\text{def}}{=} \frac{r\,l_d\nu^{-1/2} - \sum_{i=1}^{d-1} L_{di}z_i}{L_{dd}} &\leqslant \quad z_d \quad \leqslant \frac{r\,u_d\nu^{-1/2} - \sum_{i=1}^{d-1} L_{di}z_i}{L_{dd}} \stackrel{\text{def}}{=} \tilde{u}_d(r,z_1,\ldots,z_{d-1}).
\end{aligned}$$

Let $\phi(\mathbf{z};\mu,\Sigma)$ denote the density of the d-dimensional $\mathsf{N}(\mu,\Sigma)$ distribution. For the standard normal, the density is $\phi(\mathbf{z}) = \phi(\mathbf{z};\mathbf{0},I_d)$. Then, the decomposition above suggests the sequential importance sampling estimator

$$\widehat{\ell} = \frac{f_\nu(R)\phi(\mathbf{Z};\mathbf{0},I_d)}{g(R,\mathbf{Z})}$$

with $(R,\mathbf{Z})$ distributed according to the sequential importance sampling density

$$g(r,\mathbf{z}) = g(r)g(\mathbf{z}\,|\,r) = g(r)g_1(z_1\,|\,r)g_2(z_2\,|\,r,z_1)\cdots g_d(z_d\,|\,r,z_1,\ldots,z_{d-1})$$

on $\mathscr{R}$. It is then natural to choose $g(r) = f_\nu(r)$ and the truncated normal densities:

$$g_i(z_i\,|\,r,z_1,\ldots,z_{i-1}) = \frac{\phi(z_i)\mathbb{I}\{\tilde{l}_i \leqslant z_i \leqslant \tilde{u}_i\}}{\Phi(\tilde{u}_i) - \Phi(\tilde{l}_i)}, \quad i = 1,\ldots,d,$$

where Φ is the one-dimensional standard normal cdf. The estimator $\widehat{\ell}$ then simplifies to

$$\widehat{\ell}_{\text{Genz}} = \prod_{k=1}^{d}\left(\Phi(\tilde{u}_k) - \Phi(\tilde{l}_k)\right). \tag{3}$$

This estimator, proposed by Genz (2004) and discussed in Genz and Bretz (2009), is still the best method available for the estimation of (1). As we shall see in the numerical section, the variance of (3) can behave erratically, especially in the tails of the multivariate student-t distribution and in cases with strong negative correlation (as measured by $\text{Cov}(L\mathbf{Z}) = LL^\top$). For this reason, in the next section we consider an alternative importance sampling density $g(r,\mathbf{z})$ that yields a reliable and accurate estimator of ℓ in both the tails of the distribution and in the presence of negative correlation structure.

3 ESTIMATING FOR ℓ BY AN EXPONENTIALLY TILTED DISTRIBUTION

3.1 The Exponential Tilting

Instead of the Genz choice of importance sampling density described in the previous section, consider the alternative in which $g(r,\mathbf{z})$ is given as follows, where η and $\mu_1,\ldots,\mu_d$ are real-valued parameters that remain to be chosen:

$$\begin{aligned}
g(r) &= \frac{\phi(r;\eta,1)}{1-\Phi(-\eta)} = \frac{\phi(r;\eta,1)}{\Phi(\eta)}, \quad \text{for } r > 0 \\
g_k(z_k\,|\,r,z_1,\ldots,z_{k-1}) &= \frac{\phi(z_k;\mu_k,1)\mathbb{I}\{\tilde{l}_k \leqslant z_k \leqslant \tilde{u}_k\}}{\Phi(\tilde{u}_k-\mu_k) - \Phi(\tilde{l}_k-\mu_k)}, \quad \text{for } k = 1,\ldots,d.
\end{aligned} \tag{4}$$

In other words, if $\mathsf{TN}_{(a,b)}(\mu,\sigma^2)$ denotes the $\mathsf{N}(\mu,\sigma^2)$ distribution truncated to the interval (a,b), then

$$\begin{aligned} R &\sim \mathsf{TN}_{(0,\infty)}(\eta,1) \\ Z_k \,|\, R, Z_1,\ldots,Z_{k-1} &\sim \mathsf{TN}_{(\tilde{l}_k,\tilde{u}_k)}(\mu_k,1), \quad k=1,\ldots,d. \end{aligned}$$

Denoting $\mu=(\mu_1,\ldots,\mu_d)^\top$, we can write the logarithm of the likelihood ratio as

$$\begin{aligned} \psi(r,\mathbf{z};\eta,\mu) = \frac{\|\mu\|^2}{2} - \mathbf{z}^\top\mu + \frac{\eta^2}{2} - r\eta + (\nu-1)\ln r + \ln\Phi(\eta) + \text{const.} \\ + \sum_{k=1}^{d} \ln\left[\Phi(\tilde{u}_k(r,z_1,\ldots,z_{k-1})-\mu_k) - \Phi(\tilde{l}_k(r,z_1,\ldots,z_{k-1})-\mu_k)\right], \end{aligned}$$

so that $\ell=\mathbb{E}\left[\mathrm{e}^{\psi(R,\mathbf{Z};\eta,\mu)}\right]$ for $(R,\mathbf{Z})\sim g(r,\mathbf{z})$. It remains to choose the parameters η and μ so that the estimator $\widehat{\ell}_1=\mathrm{e}^{\psi(R,\mathbf{Z};\eta,\mu)}$ has a well-behaved relative error. A simple (heuristic) way of selecting (η,μ) in our setting is to minimize the worst possible behavior of the likelihood ratio $\mathrm{e}^{\psi(r,\mathbf{z};\eta,\mu)}$. In other words, we solve the optimization program

$$\inf_{\eta,\mu} \sup_{(r,\mathbf{z})\in\mathscr{R}} \psi(r,\mathbf{z};\eta,\mu). \tag{5}$$

A prime motivation for minimizing (5) is that

$$\mathrm{Var}[\widehat{\ell}_1] = \mathbb{E}\left[\exp(2\psi(R,\mathbf{Z};\eta,\mu))\right] - \ell^2 \le \exp\left[2\sup_{(r,\mathbf{z})\in\mathscr{R}} \psi(r,\mathbf{z};\eta,\mu)\right] - \ell^2,$$

and we want to select the parameter values that minimize this upper bound on the variance. Another appealing feature of (5) is that it has a unique solution that can be found by solving a convex optimization program. The idea is similar to the one described in Botev (2014), where ψ depends only on $\mathbf{z}$ and μ. Thus, we can see in retrospect that the importance function $g(r)$ and its tilting parameter η were chosen so that this convexity is preserved as shown in the following theorem, proved in Appendix A.

Theorem 3.1 (Parameter Selection)**.** *For $\nu \geqslant 1$ the saddle-point program* (5) *has a unique solution, denoted $(r^*,\mathbf{z}^*;\eta^*,\mu^*)$, which coincides with the solution of the convex optimization program:*

$$\begin{aligned} &\max_{r,\mathbf{z},\eta,\mu} \psi(r,\mathbf{z};\eta,\mu) \\ &\textit{subject to: } \partial\psi/\partial\eta = 0, \quad \partial\psi/\partial\mu = \mathbf{0}, \quad (r,\mathbf{z})\in\mathscr{R}\,. \end{aligned} \tag{6}$$

Note that without the constraint $(r,\mathbf{z})\in\mathscr{R}$, the solution of (5) is obtained by setting the gradient of ψ with respect to all of the parameters to zero. This gives the following system of nonlinear equations ($i=1,\ldots,d$ and $Z\sim\mathsf{N}(0,1)$):

$$\begin{aligned} \frac{\partial\psi}{\partial r} &= \frac{\nu-1}{r} - \eta + \sum_k \frac{u_k\phi(\tilde{u}_k-\mu_k) - l_k\phi(\tilde{l}_k-\mu_k)}{\sqrt{\nu}L_{k,k}\mathbb{P}(\tilde{l}_k \leqslant Z+\mu_k \leqslant \tilde{u}_k)} \\ \frac{\partial\psi}{\partial\eta} &= \eta - r + \frac{\phi(\eta)}{\Phi(\eta)} \\ \frac{\partial\psi}{\partial z_i} &= -\mu_i + \sum_{k=i+1}^{d}\left(\frac{L_{k,i}}{L_{k,k}} - \mathbb{I}\{i=k\}\right)\frac{\phi(\tilde{u}_k-\mu_k)-\phi(\tilde{l}_k-\mu_k)}{\mathbb{P}(\tilde{l}_k \leqslant Z+\mu_k \leqslant \tilde{u}_k)} \\ \frac{\partial\psi}{\partial\mu_i} &= \mu_i - z_i + \frac{\phi(\tilde{u}_k-\mu_k)-\phi(\tilde{l}_k-\mu_k)}{\mathbb{P}(\tilde{l}_k \leqslant Z+\mu_k \leqslant \tilde{u}_k)}. \end{aligned} \tag{7}$$

Thus, one way of solving (6) is to solve the nonlinear system (7) and verify that its solution lies in $\mathscr{R}$. If the solution lies in $\mathscr{R}$, then there is nothing else to do. This can be much faster than calling a constrained optimization solver to solve the convex program (6). However, if the solution of (7) does not lie in $\mathscr{R}$, then we must use a proper convex solver to tackle (6).

Note that we need not simulate Z_d, because the log-likelihood ratio ψ does not depend on z_d. In fact, the independence from z_d forces $\mu_d = 0$ always, reducing the dimension of the optimization (6) from $2d$ to $2(d-1)$. The proposed estimator is summarized in the following algorithm.

Algorithm 1 : Estimating ℓ.

Require: vectors $\mathbf{u},\mathbf{l}$ of dimension d and lower triangular matrix L. Sample size n.
  Solve the convex optimization program (6) to find the unique (η^*,μ^*).
  for $i = 1,\ldots,n$ **do**
    Simulate $R \sim \mathsf{TN}_{(0,\infty)}(\eta^*,1)$
    for $k = 1,\ldots,d-1$ **do**
      Simulate $Z_k \sim \mathsf{TN}_{(\tilde{l}_k,\tilde{u}_k)}(\mu_k^*,1)$
    $\mathbf{Z} \leftarrow (Z_1,\ldots,Z_{d-1},0)^\top$
    $\ell_i \leftarrow \exp(\psi(R,\mathbf{Z};\eta^*,\mu^*))$
  $\widehat{\ell} \leftarrow \frac{1}{n}\sum_{i=1}^n \ell_i$
  $\widehat{\sigma}^2 \leftarrow \frac{1}{n}\sum_{i=1}^n (\ell_i - \widehat{\ell})^2$
  return $\widehat{\ell}$ and its estimated relative error $\widehat{\sigma}/(\sqrt{n}\,\widehat{\ell})$.

If $\psi^* = \psi(r^*,\mathbf{z}^*;\eta^*,\mu^*)$ and ℓ_{L} is a lower bound to ℓ, we can bound the relative error of the estimator $\widehat{\ell}$ by

$$\frac{\sqrt{\mathrm{Var}(\widehat{\ell})}}{\ell} \leqslant \frac{1}{\sqrt{n}}\left(\frac{\psi^*}{\ell_{\mathrm{L}}} - 1\right).$$

One possibility for constructing a nontrivial lower bound on ℓ is given in Section 3.3.

3.2 Exact i.i.d. Sample From Conditional Density

In the previous algorithm, all the n samples are kept, and they are given different weights in the estimator. But if we want an exact i.i.d. sample of fixed size n (without weighting the observations) from the conditional density (2), we must proceed differently. The following algorithm does it by acceptance-rejection. It uses the fact that ψ^* yields a nontrivial upper bound to the likelihood ratio $\exp(\psi(r,\mathbf{z};\eta^*,\mu^*)) \leqslant \exp(\psi^*)$ and to the probability $\ell = \mathbb{E}[\exp(\psi(R,\mathbf{Z};\eta^*,\mu^*))] \leqslant \exp(\psi^*)$. This upper bound leads to an acceptance-rejection scheme with proposal density $g(r,\mathbf{z};\eta^*,\mu^*)$ defined via (4). The acceptance probability in this algorithm is ℓ/ψ^*.

Algorithm 2 : Exact simulation from $f(\mathbf{x})$ in (2) via acceptance-rejection.

Require: vectors $\mathbf{u},\mathbf{l}$, lower triangular L, and optimal $(r^*,\mathbf{z}^*;\eta^*,\mu^*)$.
  repeat
    Simulate $R \sim \mathsf{TN}_{(0,\infty)}(\eta^*,1)$
    for $k = 1,\ldots,d$ **do**
      Simulate $Z_k \sim \mathsf{TN}_{(\tilde{l}_k,\tilde{u}_k)}(\mu_k^*,1)$
    Simulate $E \sim \mathsf{Exp}(1)$, independently.
  until $E \geqslant \psi(r^*,\mathbf{z}^*;\eta^*,\mu^*) - \psi(R,\mathbf{Z};\eta^*,\mu^*)$
  return $\mathbf{X} \leftarrow \sqrt{\nu}\mathbf{Z}/R$

3.3 A Simple Lower Bound for ℓ

Using Jensen's inequality, it is possible to construct a simple lower bound for $\ell = \mathbb{P}(\mathbf{l} \leqslant C\mathbf{X} \leqslant \mathbf{u})$, which as we shall see in the numerical section can sometimes (but not always) be quite tight. Let $\mathbf{Y} \sim \mathbf{t}_\nu(\mathbf{0}, \Sigma)$, where $\Sigma = CC^\top$, and let h be a density on $[\mathbf{l}, \mathbf{u}] \subseteq \mathbb{R}^d$. Then, $\ell = \mathbb{P}(\mathbf{l} \leqslant \mathbf{Y} \leqslant \mathbf{u})$ and applying Jensen's inequality to the function $x^{-2/(\nu+d)}$ we obtain

$$
\begin{aligned}
(\mathbb{P}(\mathbf{l} \leqslant \mathbf{Y} \leqslant \mathbf{u}))^{-2/(\nu+d)} &= c_2^{-2/(\nu+d)} \left(\mathbb{E}_h \left[\frac{\left(1 + \frac{1}{\nu} \mathbf{Y}^\top \Sigma^{-1} \mathbf{Y}\right)^{-(\nu+d)/2}}{h(\mathbf{Y})} \right] \right)^{-2/(\nu+d)} \\
&\leqslant c_2^{-2/(\nu+d)} \mathbb{E}_h \left[\frac{1 + \frac{1}{\nu} \mathbf{Y}^\top \Sigma^{-1} \mathbf{Y}}{[h(\mathbf{Y})]^{-2/(\nu+d)}} \right],
\end{aligned}
$$

where $c_2 = c_1 / \det(\Sigma^{1/2})$ is a normalizing constant. Therefore,

$$
\begin{aligned}
(\ell/c_2)^{-2/(\nu+d)} &\leqslant \int [h(\mathbf{y})]^{\frac{2}{\nu+d}+1} \mathrm{d}\mathbf{y} + \frac{1}{\nu} \int [h(\mathbf{y})]^{\frac{2}{\nu+d}+1} (\mathbf{y}^\top \Sigma^{-1} \mathbf{y}) \, \mathrm{d}\mathbf{y} \\
&\leqslant \int [h(\mathbf{y})]^{\frac{2}{\nu+d}+1} \mathrm{d}\mathbf{y} \left(1 + \frac{1}{\nu} \mathbb{E}_q [\mathbf{Y}^\top \Sigma^{-1} \mathbf{Y}] \right),
\end{aligned}
$$

where the density q is defined via h through

$$
q(\mathbf{y}) = \frac{[h(\mathbf{y})]^{\frac{2}{\nu+d}+1}}{\int [h(\mathbf{y})]^{\frac{2}{\nu+d}+1} \mathrm{d}\mathbf{y}},
$$

so that

$$
\int [h(\mathbf{y})]^{\frac{2}{\nu+d}+1} \mathrm{d}\mathbf{y} = \left(\int [q(\mathbf{y})]^{\frac{\nu+d}{\nu+d+2}} \mathrm{d}\mathbf{y} \right)^{-(\nu+d+2)/(\nu+d)}.
$$

Rearranging the last inequality then yields

$$
\ell \;\geqslant\; c_2 \left(\int [q(\mathbf{y})]^{\frac{\nu+d}{\nu+d+2}} \mathrm{d}\mathbf{y} \right)^{(\nu+d+2)/2} \left(1 + \frac{1}{\nu} \mathrm{tr}(\Sigma^{-1} \mathrm{Var}_q(\mathbf{Y})) + \frac{1}{\nu} \mathbb{E}_q[\mathbf{Y}]^\top \Sigma^{-1} \mathbb{E}_q[\mathbf{Y}] \right)^{-(\nu+d)/2} \tag{8}
$$

All terms on the right-hand side of (8) can be computed analytically if we choose the product form $q(\mathbf{y}) = \prod_k q_k(y_k)$, where q_k is the density of the univariate student-t distribution truncated to the interval $[l_k, u_k]$ and with ν_k degrees of freedom, location μ_k, and scale σ_k. The exact analytical expressions for the right-hand side of (8) are given in Appendix B.

The best lower bound is obtained by maximizing the right-hand side of (8) with respect to $\{\nu_i, \mu_i, \sigma_i,\ i = 1, \ldots, d\}$. This is the lower bound we use in the numerical experiments in Section 4.

4 A NUMERICAL STUDY

In this section we compare the numerical performance of our estimator with that of Genz.

In all examples the computing time to find the optimal tilting parameter (η^*, μ^*) was insignificant compared to the time it took to evaluate the n iid replications in the 'for' loop of Algorithm 1. One reason for this was that the solution of the nonlinear system (7) always belonged to the set $\mathscr{R}$ and was thus identical to the solution of the program (6), obviating the need for a convex optimization routine. For this reason, we only report the relative error of the estimators in our comparison. Note that although in general there are many ways of decomposing $\Sigma = CC^\top$, the proposed methods do not depend on the choice of C for a given Σ.

Example 4.1 (Negative Correlation)**.** Consider estimating ℓ, where $CC^\top = \Sigma$ is defined via the precision matrix (Fernández, Ferrari, and Grynberg 2007):

$$\Sigma^{-1} = \frac{1}{2} I_d + \frac{1}{2}\mathbf{1}\mathbf{1}^\top$$

In the following Table 1 we list the estimates from both methods in columns three and four with their estimated relative variances in bold font. In addition, we list: a) the best lower bound from (8) in column two; b) the upper bound ψ^* in column five; and c) the estimated acceptance probability $\widehat{\ell}/\psi^*$ in column six.

Table 1: Estimates of ℓ for $[\mathbf{l},\mathbf{u}] = [-1,\infty]^d$ with $\nu = 10$ using $n = 10^5$ replications.

d	lower bound	$\widehat{\ell}_{\text{Genz}}$	$\widehat{\ell}$	ψ^*	accept. prob.
5	0.15	0.197 **(0.21%)**	0.197 **(0.18%)**	0.33	59%
10	0.013	0.032 **(0.49%)**	0.032 **(0.20%)**	0.063	50%
20	1.16×10^{-4}	0.00161 **(1.8%)**	0.00163 **(0.23%)**	0.00385	42%
30	1.24×10^{-6}	1.53×10^{-4} **(2.8%)**	1.51×10^{-4} **(0.26%)**	3.92×10^{-4}	38%
40	1.54×10^{-8}	1.81×10^{-5} **(5.4%)**	2.08×10^{-5} **(0.29%)**	5.68×10^{-5}	36%
50	2.17×10^{-10}	3.63×10^{-6} **(15%)**	3.74×10^{-6} **(0.25%)**	1.06×10^{-5}	35%
100	3.35×10^{-19}	3.44×10^{-9} **(51%)**	6.99×10^{-9} **(0.28%)**	2.11×10^{-8}	33%
150	1.29×10^{-27}	6.35×10^{-11} **(47%)**	9.29×10^{-11} **(0.27%)**	2.85×10^{-10}	32%

From the table we can conclude the following. First, the lower bound (8) is not useful in this example. From a range of simulations we found that the bound is typically tight only when we consider tail-like regions such as $[\gamma,\infty]^d$ for $\gamma > 0$, which is not the case here. Second, as d increases the performance of the Genz estimator rapidly deteriorates. In contrast, the relative error of $\widehat{\ell}$ remains stable for all d. The acceptance probability in column six indicates that Algorithm 3.2 is useful for simulating from the conditional density. Note that a naive acceptance-rejection scheme in which we simulate $\mathbf{X} \sim \mathbf{t}_\nu$ until $\mathbf{l} \leqslant C\mathbf{X} \leqslant \mathbf{u}$ is only practical up to about dimension $d = 30$, beyond which the acceptance probability ℓ is too small.

Now, consider the same setting, but this time with the orthant region $[\mathbf{l},\mathbf{u}] = [0,\infty]^d$.

Table 2: Estimates of ℓ for $[\mathbf{l},\mathbf{u}] = [0,\infty]^d$ with $\nu = 10$ using $n = 10^5$ replications.

d	lower bound	$\widehat{\ell}_{\text{Genz}}$	$\widehat{\ell}$	ψ^*	accept. prob.
5	0.00190	0.00193 **(0.39%)**	0.00192 **(0.15%)**	0.0030	63%
10	1.55×10^{-7}	1.69×10^{-7} **(2.2%)**	1.58×10^{-7} **(0.16%)**	2.67×10^{-7}	59%
20	2.76×10^{-17}	1.18×10^{-17} **(43%)**	2.98×10^{-17} **(0.16%)**	5.34×10^{-17}	55%
30	3.29×10^{-28}	1.29×10^{-33} **(98%)**	3.79×10^{-28} **(0.13%)**	6.99×10^{-28}	54%
40	6.89×10^{-40}	–	8.48×10^{-40} **(0.15%)**	1.58×10^{-39}	53%
50	4.00×10^{-52}	–	5.23×10^{-52} **(0.21%)**	9.91×10^{-52}	52%
100	1.02×10^{-118}	–	1.71×10^{-118} **(0.19%)**	3.33×10^{-118}	51%
150	5.18×10^{-191}	–	1.03×10^{-190} **(0.30%)**	2.02×10^{-190}	50%

The results in the table above indicate that the lower bound is now useful. Another interesting point is that the performance of the Genz estimator now degrades much more rapidly and fails to give meaningful estimates for $d > 20$.

Although not displayed here, the effect of the exponential tilting is even more dramatic with the tail-like region $[1,\infty]^d$. In fact, we conjecture that the proposed estimator exhibits *bounded relative error* as $\gamma \uparrow \infty$ when $\ell(\gamma) = \mathbb{P}(C\mathbf{X} \geqslant \gamma CC^\top \mathbf{l}^*)$, where $\mathbf{l}^* > \mathbf{0}$. This would mean that $\limsup_{\gamma\uparrow\infty} \mathrm{Var}(\widehat{\ell})/\ell^2 < \infty$ (L'Ecuyer et al. 2010). See also Asmussen and Glynn (2007) and Kroese et al. (2011), Chapter 10 for discussions of efficiency measures when estimating rare-event probabilities.

Example 4.2 (Positive Correlation)**.** Consider the case $[\mathbf{l},\mathbf{u}] = [1,2]^d$ with

$$\Sigma = CC^\top = (1-\rho)I_d + \rho\mathbf{1}\mathbf{1}^\top,$$

where $\rho = 0.95$. The table below displays the results, which suggest that in cases with strong positive correlation, the estimator $\widehat{\ell}_{\text{Genz}}$ is more accurate and reliable. Further, we observed that the improvement due to exponential tilting in such cases is marginal and the lower bound (8) is not tight.

Table 3: Estimates of ℓ for $[\mathbf{l},\mathbf{u}] = [1,2]^d$ with $\nu = 10$ using $n = 10^5$ replications.

d	lower bound	$\widehat{\ell}_{\text{Genz}}$	$\widehat{\ell}$	ψ^*	accept. prob.
5	0.046	0.099 **(0.19%)**	0.099 **(0.21%)**	0.22	44%
30	0.010	0.060 **(0.27%)**	0.060 **(0.29%)**	0.20	29%
50	0.0059	0.0520 **(0.38%)**	0.0518 **(0.46%)**	0.20	25%
100	0.0022	0.0424 **(0.46%)**	0.0424 **(0.35%)**	0.19	22%
150	0.0012	0.037 **(0.59%)**	0.037 **(0.49%)**	0.18	20%

Example 4.3 (Random Covariance Matrix)**.** In this example we consider test cases in which $\Sigma = CC^\top$ is a random draw from a large sample space of possible covariance matrices. A popular method for simulating random positive-definite test matrices is that of Davies and Higham (2000), who simulate correlation matrices with eigenvalues uniformly distributed over the simplex $\{\lambda : \sum_i \lambda_i = d,\ \lambda_i > 0\}$. Table 4 and Figure 4.3 below show the five-number summary and boxplots of the empirical distributions of the relative errors of estimators $\widehat{\ell}$ and $\widehat{\ell}_{\text{Genz}}$ based on 100 independent trials (100 replications of the entire scheme with a sample size n each). For each trial we simulated a different (random) scale matrix Σ according to the mechanism of Davies and Higham (2000). In this example we set $[\mathbf{l},\mathbf{u}] = [1,\infty]^{100}$ and for each of the 100 independent trials we used $n = 10^5$.

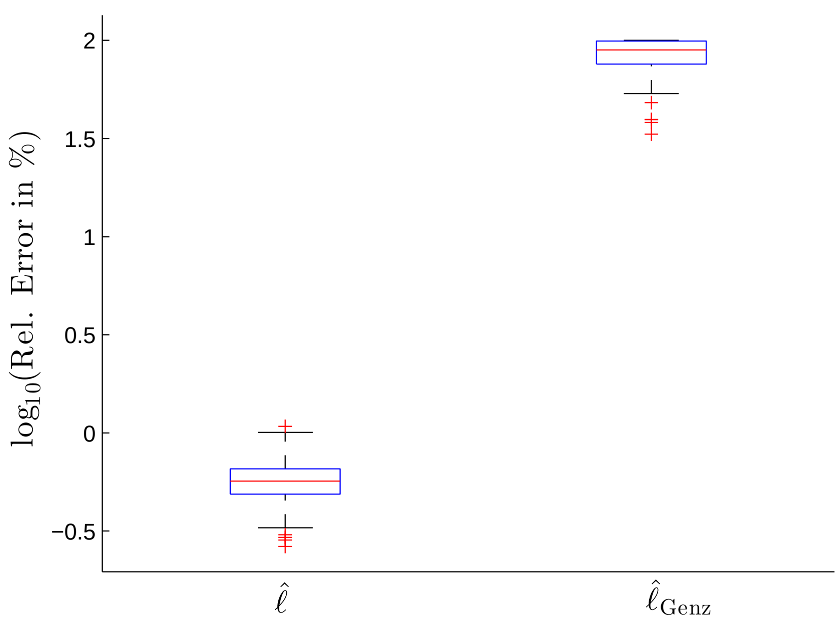

Figure 1: Empirical distribution of relative errors of $\widehat{\ell}$ and $\widehat{\ell}_{\text{Genz}}$ when $[\mathbf{l},\mathbf{u}] = [1,\infty]^{100}$ and $n = 10^5$.

Table 4: A five number summary of the distributions of the relative errors.

	min	1-st quartile	median	3-rd quartile	max
rel. error of $\widehat{\ell}$	0.26%	0.48%	0.56%	0.65%	1.08%
rel. error of $\widehat{\ell}_{\text{Genz}}$	33%	75%	89%	99%	100%
$\psi^*/\widehat{\ell}$	16	46	66	100	470

It is clear that $\widehat{\ell}_{\text{Genz}}$ is not a useful estimator in this setting, because in the best of cases it could only manage a relative error of about 30%. The last row of Table 4 displays the average number of trials needed before acceptance in Algorithm 3.2.

For a more challenging example suppose each element of matrix C is Cauchy distributed with location 0 and scale 0.01^2. In other words, $C_{i,j} \stackrel{\text{iid}}{\sim} \mathbf{t}_1(0, 0.01^2)$ and $\Sigma = CC^\top$. Here we consider the case $[\mathbf{l}, \mathbf{u}] = [0, \infty]^{100}$. The following table and graph display the empirical distributions of the relative errors of $\widehat{\ell}$ and $\widehat{\ell}_{\text{Genz}}$ based on 1000 independent replications of the experiment.

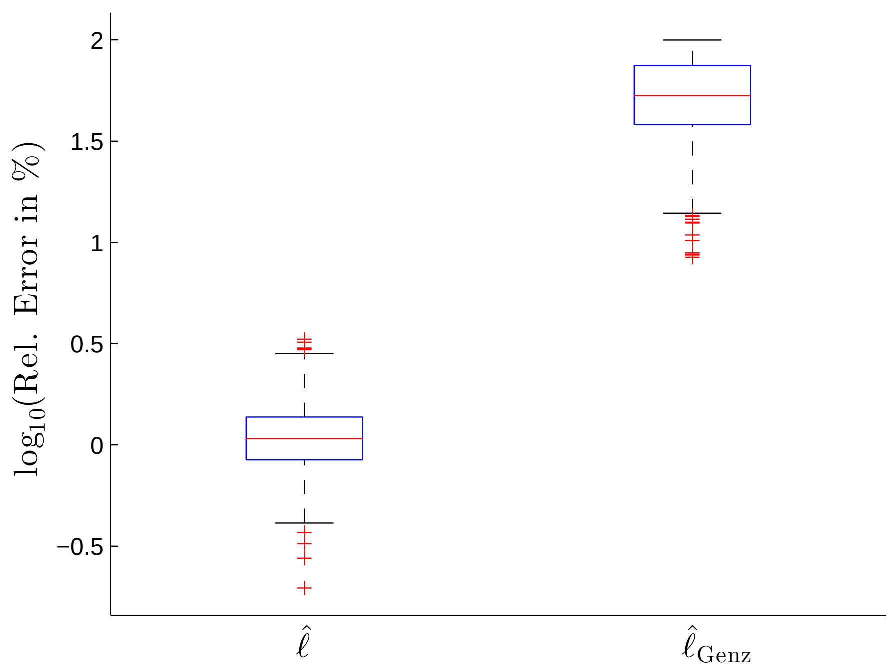

Figure 2: Empirical distribution of relative errors of $\widehat{\ell}$ and $\widehat{\ell}_{\text{Genz}}$ when $[\mathbf{l}, \mathbf{u}] = [0, \infty]^{100}$ and $n = 10^5$.

Table 5: A five number summary of the distributions of the relative errors.

	min	1-st quartile	median	3-rd quartile	max
rel. error of $\widehat{\ell}$	0.20%	0.84%	1.0%	1.4%	3.33%
rel. error of $\widehat{\ell}_{\text{Genz}}$	8.4%	38%	53%	75%	99.9%
$\psi^*/\widehat{\ell}$	80	500	900	1600	14000

5 CONCLUDING REMARKS

We have presented a new method for simulation from the truncated multivariate student-t distribution and estimation of the normalizing constant of the associated truncated density. The method combines exponential tilting with convex optimization. Numerical experiments suggest that the method is effective in many different settings and not just in the tails of the distribution. The numerical results also suggest that the approach yields insignificant improvement over the Genz estimator when we have strong positive correlation. One reason for this seems to be that the Genz estimator already works quite well in such settings, making it difficult to improve upon. At the same time the Genz estimator performs extremely poorly in the absence of positive correlation structure. All of these observations invite further theoretical study. For example, it would be interesting to see if an efficiency result, such as vanishing relative error, can be established in an appropriate tail asymptotic regime. We intend to investigate these issues in upcoming work.

APPENDIX A: PROOF OF THEOREM 1

We show that ψ is a concave function of vector $(r,\mathbf{z})$ and a convex function of (η,μ). To this end, recall that if $f:\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\to\mathbb{R}$ is a log-concave function, then the marginal

$$g(\mathbf{x})=\int_{\mathbb{R}^{d_2}} f(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y}$$

is a log-concave function as well, see Prékopa (1973). Also recall that the indicator function $\mathbb{I}\{\mathbf{x}\in\mathscr{C}\}$ of a convex set $\mathscr{C}$ is a log-concave function of $\mathbf{x}$ and that the product of log-concave functions is again log-concave. We can write

$$\mathbb{I}\left\{L_{kk}(x+\mu_k)+\sum_{i=1}^{k-1}L_{ki}z_i-\frac{rl_k}{\sqrt{\nu}}\geqslant 0\right\}=\mathbb{I}\{(r,\mathbf{z})\in\mathscr{C}_1\}$$

and

$$\mathbb{I}\left\{L_{kk}(x+\mu_k)+\sum_{i=1}^{k-1}L_{ki}z_i-\frac{ru_k}{\sqrt{\nu}}\leqslant 0\right\}=\mathbb{I}\{(r,\mathbf{z})\in\mathscr{C}_2\}$$

for some convex sets $\mathscr{C}_1$ and $\mathscr{C}_2$. It follows that

$$\ln\left[\Phi(\tilde{u}_k-\mu_k)-\Phi(\tilde{l}_k-\mu_k)\right]=\ln\int\phi(x;0,1)\times\mathbb{I}\{(r,\mathbf{z})\in\mathscr{C}_1\}\times\mathbb{I}\{(r,\mathbf{z})\in\mathscr{C}_2\}\mathrm{d}x$$

is concave in $(r,\mathbf{z})$ by Prékopa's result. Since $\ln(r)$ is concave and the sum of concave functions $\sum_k\ln(\Phi(\tilde{u}_k-\mu_k)-\Phi(\tilde{l}_k-\mu_k))$ is concave, it follows that ψ is concave in $(r,\mathbf{z})$. Next, note that $\frac{1}{2}\eta^2-r\eta+\ln\Phi(\eta)$ is convex in η, because (up to a normalizing constant)

$$\frac{1}{2}\eta^2+\ln\Phi(\eta)=\ln\int_{-\infty}^{0}\phi(x)\exp(-x\eta)\mathrm{d}x+\text{const.}$$

is the cumulant function of the normal distribution, truncated to the interval $(-\infty,0]$. A similar reasoning shows that

$$\frac{1}{2}\mu_k^2-z_k\mu_k+\ln\int_{\tilde{l}_l-\mu_k}^{\tilde{u}_k-\mu_k}\phi(x)\mathrm{d}x$$

is convex in μ_k and since a sum of convex functions is convex, ψ is convex in the vector (η,μ). Thus, the concave-convex function $\psi(r,\mathbf{z};\eta,\mu)$ satisfies the saddle-point condition $\inf_{\eta,\mu}\sup_{r,\mathbf{z}}\psi(r,\mathbf{z};\eta,\mu)=\sup_{r,\mathbf{z}}\inf_{\eta,\mu}\psi(r,\mathbf{z};\eta,\mu)$. Recall that if for each $\mathbf{y}$ the function $f(\mathbf{x},\mathbf{y})$ is convex, then the pointwise supremum $\sup_{\mathbf{y}}f(\mathbf{x},\mathbf{y})$ is also convex. Therefore, $\inf_{\eta,\mu}\sup_{r,\mathbf{z}}\psi(r,\mathbf{z};\eta,\mu)$ has the same value as the concave optimization $\sup_{r,\mathbf{z}}\psi(r,\mathbf{z};\eta,\mu)$ subject to the gradient of ψ with respect to (η,μ) being equal to the zero vector. Imposing the restriction $(r,\mathbf{z})\in\mathscr{R}$, where $\mathscr{R}$ is a convex set, does not change the argument, which leads us to the optimization problem (6).

6 APPENDIX B: ANALYTICAL EXPRESSIONS FOR LOWER BOUND

The right-hand side of (8) is available analytically as follows. Let

$$t_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}(1+x^2/\nu_i)^{-(\nu+1)/2}$$

denote the density of univariate student-t distribution with ν degrees of freedom, and let

$$T_\nu(x) = 1-\frac{1}{2}\frac{1}{B(\nu/2,1/2)}\int_0^{\nu/(\nu+x^2)} t^{\frac{\nu}{2}-1}(1-t)^{-\frac{1}{2}}\mathrm{d}t$$

be the corresponding cdf. Then, we can define and compute the following quantities:

$$\alpha_i \stackrel{\text{def}}{=} \frac{l_i-\mu_i}{\sigma_i}, \quad \beta_i \stackrel{\text{def}}{=} \frac{u_i-\mu_i}{\sigma_i}, \quad c_i \stackrel{\text{def}}{=} T_{\nu_i}(\beta_i)-T_{\nu_i}(\alpha_i)$$

$$q_i(x_i) \stackrel{\text{def}}{=} \frac{\frac{1}{\sigma_i}t_{\nu_i}\left(\frac{x_i-\mu_i}{\sigma_i}\right)}{c_i}, \qquad x_i\in[l_i,u_i]$$

$$\frac{\mathbb{E}_{q_i}[X_i]-\mu_i}{\sigma_i} = \frac{1}{c_i}\frac{\nu_i}{\nu_i-1}(t_{\nu_i}(\alpha_i)-t_{\nu_i}(\beta_i)) + \frac{1}{c_i}\frac{1}{\nu_i-1}(t_{\nu_i}(\alpha_i)\alpha_i^2-t_{\nu_i}(\beta_i)\beta_i^2)$$

$$\mathbb{E}_{q_i}\left[\frac{X_i-\mu_i}{\sigma_i}\right]^2 = \frac{(\nu_i-1)\nu_i}{\nu_i-2}\,\frac{T_{\nu_i-2}(\beta_i\sqrt{(\nu_i-2)/\nu_i})-T_{\nu_i-2}(\alpha_i\sqrt{(\nu_i-2)/\nu_i})}{c_i}-\nu_i$$

These calculations give us all quantities on the right-hand size of (8), except for $\int[q(\mathbf{y})]^{\frac{\nu+d}{\nu+d+2}}\mathrm{d}\mathbf{y} = \prod_i\int[q_i(y_i)]^{\frac{\nu+d}{\nu+d+2}}\mathrm{d}y_k$. To compute the last integral let $\xi_i = \frac{(\nu_i+1)(\nu+d)}{\nu+d+2}-1$, and then use the following analytical expressions:

$$\int_{l_i}^{u_i}[q_i(x)]^{\frac{\nu+d}{\nu+d+2}}\mathrm{d}x = \left(\frac{\Gamma((\nu_i+1)/2)}{\Gamma(\nu_i/2)\sqrt{\nu_i\pi}\sigma_i c_i}\right)^{\frac{\nu+d}{\nu+d+2}}\int_{l_i}^{u_i}\left(1+\frac{(x-\mu_i)^2}{\nu_i\sigma_i^2}\right)^{-\frac{\nu_i+1}{2}\frac{\nu+d}{\nu+d+2}}\mathrm{d}x,$$

where $\frac{\nu_i+1}{2}\frac{\nu+d}{\nu+d+2}=\frac{\xi_i+1}{2}$ and

$$\int_{l_i}^{u_i}\left(1+\frac{(x-\mu_i)^2}{\nu_i\sigma_i^2}\right)^{-\frac{\xi_i+1}{2}}\mathrm{d}x = \sigma_i\sqrt{\nu/\xi_i}\frac{\sqrt{\xi_i\pi}\,\Gamma(\xi_i/2)}{\Gamma((\xi_i+1)/2)}\left(T_{\xi_i}(\beta_i\sqrt{\xi_i/\nu})-T_{\xi_i}(\alpha_i\sqrt{\xi_i/\nu})\right).$$

ACKNOWLEDGMENTS

Zdravko Botev has been supported by the *Australian Research Council Discovery Early Career Researcher Award* DE140100993 and the *Early Career Researcher Grant* of the School of Mathematics and Statistics at the University of New South Wales (UNSW), Sydney, Australia. Pierre L'Ecuyer received support from an NSERC-Canada Discovery Grant, a Canada Research Chair, and an Inria International Chair.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation*. Springer-Verlag.

Z. I. Botev 2014. "The Normal Law Under Linear Restrictions: Simulation and Estimation via Minimax Tilting". Submitted.

Bucklew, J. A. 2004. *Introduction to Rare Event Simulation*. New York: Springer-Verlag.

Davies, P. I., and N. J. Higham. 2000. "Numerically stable generation of correlation matrices and their factors". *BIT Numerical Mathematics* 40 (4): 640–651.

Fernández, P. J., P. A. Ferrari, and S. P. Grynberg. 2007. "Perfectly random sampling of truncated multinormal distributions". *Advances in Applied Probability* 39:973–990.

Genz, A. 2004. "Numerical computation of rectangular bivariate and trivariate normal and t probabilities". *Statistics and Computing* 14 (3): 251–260.

Genz, A., and F. Bretz. 2002. "Comparison of methods for the computation of multivariate t probabilities". *Journal of Computational and Graphical Statistics* 11 (4): 950–971.

Genz, A., and F. Bretz. 2009. *Computation of multivariate normal and t probabilities*, Volume 195 of *Lecture Notes in Statistics*. Springer-Verlag.

Kroese, D. P., T. Taimre, and Z. I. Botev. 2011. *Handbook of Monte Carlo Methods*. John Wiley & Sons.

L'Ecuyer, P., J. H. Blanchet, B. Tuffin, and P. W. Glynn. 2010. "Asymptotic Robustness of Estimators in Rare-Event Simulation". *ACM Transactions on Modeling and Computer Simulation* 20 (1): Article 6.

Prékopa, A. 1973. "On logarithmic concave measures and functions". *Acta Scientiarum Mathematicarum* 34:335–343.

Yu, J.-W., and G.-L. Tian. 2011. "Efficient algorithms for generating truncated multivariate normal distributions". *Acta Mathematicae Applicatae Sinica, English Series* 27 (4): 601–612.

AUTHOR BIOGRAPHIES

ZDRAVKO I. BOTEV is a 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. He has written jointly with D. P. Kroese and T. Taimre a *Handbook of Monte Carlo Methods* published by John Wiley & Sons in 2011.

PIERRE L'ECUYER is Professor in the Département d'Informatique et de Recherche Opérationnelle, at the Université de Montréal, Canada. He holds the Canada Research Chair in Stochastic Simulation and Optimization and an Inria International Chair in Rennes, France. He is a member of the CIRRELT and GERAD research centers. His main research interests are random number generation, quasi-Monte Carlo methods, efficiency improvement via variance reduction, sensitivity analysis and optimization of discrete-event stochastic systems, and discrete-event simulation in general. He has served as Editor-in-Chief for *ACM Transactions on Modeling and Computer Simulation* from 2010 to 2013. He is currently Associate Editor for *ACM Transactions on Mathematical Software*, *Statistics and Computing*, and *International Transactions in Operational Research*. More information can be found on his web page: http://www.iro.umontreal.ca/~lecuyer.