ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)

The normal law under linear restrictions: simulation and estimation via minimax tilting

Z. I. Botev

University of New South Wales, Sydney, Australia

[Received November 2014. Final revision November 2015]

Summary. Simulation from the truncated multivariate normal distribution in high dimensions is a recurrent problem in statistical computing and is typically only feasible by using approximate Markov chain Monte Carlo sampling. We propose a minimax tilting method for exact independently and identically distributed data simulation from the truncated multivariate normal distribution. The new methodology provides both a method for simulation and an efficient estimator to hitherto intractable Gaussian integrals. We prove that the estimator has a rare vanishing relative error asymptotic property. Numerical experiments suggest that the scheme proposed is accurate in a wide range of set-ups for which competing estimation schemes fail. We give an application to exact independently and identically distributed data simulation from the Bayesian posterior of the probit regression model.

Keywords: Exact simulation; Exponential tilting; Linear inequalities; Multivariate normal distribution; Polytope probabilities; Probit posterior simulation

1. Introduction

More than a century ago Galton (1889) observed that he scarcely knew

> 'anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the law of frequency of error. The law would have been personified by the Greeks if they had known of it'.

In this paper we address some hitherto intractable computational problems that are related to the d-dimensional multivariate normal law under linear restrictions:

$$f(\mathbf{z})=\frac{1}{l}\exp\left(-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{z}\right)\mathbb{I}\,(\mathbf{l}\leqslant A\mathbf{z}\leqslant\mathbf{u}),\qquad \mathbf{z}=(z_1,\ldots,z_d)^{\mathrm{T}},\quad A\in\mathbb{R}^{m\times d},\quad \mathbf{u},\mathbf{l}\in\mathbb{R}^{m},\qquad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathrm{rank}(A)=m\leqslant d$ and $l=\mathbb{P}(\mathbf{l}\leqslant A\mathbf{Z}\leqslant\mathbf{u})$ is the probability that a random vector $\mathbf{Z}$ with standard normal distribution in d-dimensions (i.e. $\mathbf{Z}\sim N(\mathbf{0},I_d)$) falls in the H-polytope defined by the linear inequalities.

Aesthetic considerations aside, the problem of estimating l or simulating from $f(\mathbf{z})$ arises frequently in various contexts such as Markov random fields (Bolin and Lindgren, 2015), inference for spatial processes (Wadsworth and Tawn, 2014), likelihood estimation for max-stable processes (Huser and Davison, 2013; Genton *et al.*, 2011), computation of simultaneous confidence bands (Azaïs *et al.*, 2010), uncertainty regions for latent Gaussian models (Bolin and

Address for correspondence: Z. I. Botev, Department of Statistics, University of New South Wales, High Street Kensington, Sydney, NSW 2052, Australia.
E-mail: botev@unsw.edu.au

1369–7412/16/79000

Lindgren, 2015), fitting mixed effects models with censored data (Grün and Hornik, 2012) and probit regression (Albert and Chib, 1993), to name a few.

For the reasons that were outlined above, the problem of estimating l accurately has received considerable attention. For example, Craig (2008), Miwa *et al.* (2003), Gassmann (2003), Genz (2004), Hayter and Lin (2012, 2013) and Nomura (2016) considered approximation methods for special cases (orthant, bivariate or trivariate probabilities) and Geweke (1991), Genz (1992), Joe (1995), Vijverberg (1997), Sándor and András (2004) and Nomura (2014) considered estimation schemes that are applicable for general l. Extensive comparisons between the numerous proposals in the literature (Genz and Bretz, 2002, 2009; Gassmann *et al.*, 2002) indicate that the method of Genz (1992) is the most accurate across a wide range of test problems of medium and large dimensions. Even in low dimensions ($d \leqslant 7$), the method compares favourably with highly specialized routines for orthant probabilities (Miwa *et al.*, 2003; Craig, 2008). For this reason, Genz's method is the default choice across different software platforms like Fortran, MATLAB® and R.

One of the goals of this paper is to propose a new methodology, which not only yields an unbiased estimator that is orders of magnitude less variable than the Genz estimator but also works reliably in cases where the Genz estimator and other alternatives fail to deliver meaningful estimates (e.g. relative errors close to 100%). (MATLAB® and R implementations are available from MATLAB® CENTRAL, `http://www.mathworks.com/matlabcentral/fileexchange/53796-truncated-multivariate-normal`, and the Comprehensive R Archive Network repository (under the name `TruncatedNormal`), as well as from the author's Web site: `http://web.maths.unsw.edu.au/~zdravkobotev/`.)

The obverse to the problem of estimating l is simulation from the truncated multivariate normal $f(\mathbf{z})$. Despite the close relationship between the two problems, they have rarely been studied concurrently (Botts, 2013; Chopin, 2011; Fernández *et al.*, 2007; Philippe and Robert, 2003). Thus, another goal of this paper is to provide an exact accept–reject sampling scheme for simulation from $f(\mathbf{z})$ in high dimensions, which traditionally calls for approximate Markov chain Monte Carlo simulation. Such a scheme can either obviate the need for Gibbs sampling (Fernández *et al.*, 2007) or can be used to accelerate Gibbs sampling through the blocking of hundreds of highly dependent variables (Chopin, 2011). Unlike existing algorithms, the accept–reject sampler that is proposed in this paper enjoys high acceptance rates in over 100 dimensions and takes about the same time as one cycle of Gibbs sampling.

The gist of the method is to find an exponential tilting of a suitable importance sampling measure by solving a minimax (saddle point) optimization problem. The optimization can be solved efficiently, because it exploits log-concavity properties of the normal distribution. The method permits us to construct an estimator with a tight deterministic bound on its relative error and a concomitant exact stochastic confidence interval. Our importance sampling proposal builds on the celebrated Genz construction, but the addition of minimax tilting ensures that the new estimator enjoys theoretically better variance properties than the Genz estimator. In an appropriate asymptotic tail regime, minimax tilting yields an estimator with the vanishing relative error (VRE) property (Kroese *et al.*, 2011). Within the light-tailed exponential family, Monte Carlo estimators rarely have the valuable VRE property (L'Ecuyer *et al.*, 2010) and so far no estimator of l with such properties has been proposed. The VRE property implies, for example, that the new accept–reject instrumental density converges in total variation to the target density $f(\mathbf{z})$, rendering sampling in the tails of the truncated normal distribution asymptotically feasible. In this paper we focus on the multivariate normal law because of its central position in statistics, but the methodology proposed can be easily generalized to other multivariate elliptic distributions.

2. Background on separation of variables estimator

We first briefly describe the separation-of-variables (SOV) estimator of Genz (1992) (see also Geweke (1991)). Let $A = LQ^{\mathrm{T}}$ be the LQ-decomposition of the matrix A, where L is $m \times d$ lower triangular with non-negative entries down the main diagonal and $Q^{\mathrm{T}} = Q^{-1}$ is $d \times d$ orthonormal. A simple change of variable $\mathbf{x} \leftarrow Q^{\mathrm{T}}\mathbf{z}$ then yields

$$l = \mathbb{P}(\mathbf{l} \leqslant L\mathbf{Z} \leqslant \mathbf{u}) = \int_{\mathbf{l} \leqslant L\mathbf{x} \leqslant \mathbf{u}} \phi(\mathbf{x}; \mathbf{0}, I)\,\mathrm{d}\mathbf{x},$$

where $\phi(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ denotes the probability density function (PDF) of the $N(\boldsymbol{\mu}, \Sigma)$ distribution. For simplicity of notation, we henceforth assume that $m = d$ so that L is full rank. The case of $m < d$ is considered later in Section 5. Genz (1992) decomposed the region $\mathscr{C} = \{\mathbf{x} : \mathbf{l} \leqslant L\mathbf{x} \leqslant \mathbf{u}\}$ sequentially as follows:

$$\begin{aligned}
\tilde{l}_1 &\overset{\text{def}}{=} \frac{l_1}{L_{11}} \leqslant x_1 \leqslant \frac{u_1}{L_{11}} \overset{\text{def}}{=} \tilde{u}_1, \\
\tilde{l}_2(x_1) &\overset{\text{def}}{=} \frac{l_2 - L_{21}x_1}{L_{22}} \leqslant x_2 \leqslant \frac{u_2 - L_{21}x_1}{L_{22}} \overset{\text{def}}{=} \tilde{u}_2(x_1), \\
&\vdots \\
\tilde{l}_d(x_1, \ldots, x_{d-1}) &\overset{\text{def}}{=} \frac{l_d - \sum_{j=1}^{d-1} L_{dj}x_j}{L_{dd}} \leqslant x_d \leqslant \frac{u_d - \sum_{j=1}^{d-1} L_{dj}x_j}{L_{dd}} \overset{\text{def}}{=} \tilde{u}_d(x_1, \ldots, x_{d-1}).
\end{aligned}$$

This decomposition motivates the SOV estimator of l,

$$\hat{l} = \frac{\phi(\mathbf{X}; \mathbf{0}, I)}{g(\mathbf{X})}, \qquad \mathbf{X} \sim g(\mathbf{x}), \tag{2}$$

where g is an importance sampling density over the set $\mathscr{C}$ and in the SOV form

$$g(\mathbf{x}) = g_1(x_1)g_2(x_2|x_1)\ldots g_d(x_d|x_1, \ldots, x_{d-1}), \qquad \mathbf{x} \in \mathscr{C}. \tag{3}$$

We denote the measure corresponding to g by $\mathbb{P}_{\mathbf{0}}$. The Genz SOV estimator, which we denote by $\mathring{l}$ to distinguish it from the more general $\hat{l}$, is obtained by selecting for all $k = 1, \ldots, d$

$$g_k(x_k|x_1, \ldots, x_{k-1}) \propto \phi(x_k; 0, 1)\mathbb{I}(\tilde{l}_k \leqslant x_k \leqslant \tilde{u}_k). \tag{4}$$

Denoting by $\Phi(\cdot)$ the cumulative density function of the standard normal distribution, this gives the algorithm in Table 1.

Table 1. Algorithm 1 (SOV estimator)

```
Require the lower triangular $L$ such that $A = LQ^{\mathrm{T}}$, bounds $\mathbf{l}, \mathbf{u}$, and uniform sequence $U_1, \ldots, U_{d-1} \overset{\text{IID}}{\sim} U(0,1)$;
  for $k = 1, 2, \ldots, d-1$ do
    simulate $X_k \sim N(0,1)$ conditionally on $\tilde{l}_k(X_1, \ldots, X_{k-1}) \leqslant X_k \leqslant \tilde{u}_k(X_1, \ldots, X_{k-1})$ using the inverse transform
      method, i.e. set
          $X_k = \Phi^{-1}[\Phi(\tilde{l}_k) + U_k\{\Phi(\tilde{u}_k) - \Phi(\tilde{l}_k)\}]$
return $\mathring{l} = \prod_{k=1}^{d}[\Phi\{\tilde{u}_k(X_1, \ldots, X_{k-1})\} - \Phi\{\tilde{l}_k(X_1, \ldots, X_{k-1})\}]$
```

Table 2. Algorithm 2 (accept–reject simulation from f)

```
*Require* the supremum of likelihood ratio $c = \sup_{\mathbf{x} \in \mathscr{C}} \phi(\mathbf{x}; \mathbf{0}, I)/g(\mathbf{x})$
Simulate $U \sim U(0,1)$ and $\mathbf{X} \sim g(\mathbf{x})$, independently:
*while* $cU > \phi(\mathbf{X}; \mathbf{0}, I)/g(\mathbf{X})$ *do*
    simulate $U \sim U(0,1)$ and $\mathbf{X} \sim g(\mathbf{x})$, independently;
*return* $\mathbf{X}$, an outcome from the truncated multivariate normal density
    $f$ in expression (1)
```

Algorithm 1 can be repeated n times to obtain the independently and identically distributed (IID) sample $\mathring{l}_1, \ldots, \mathring{l}_n$ that is used for the construction of the unbiased point estimator $\bar{l} = (\mathring{l}_1 + \ldots + \mathring{l}_n)/n$ and its approximate 95% confidence interval $(\bar{l} \pm 1.96S/\sqrt{n})$, where S is the sample standard deviation of $\mathring{l}_1, \ldots, \mathring{l}_n$.

2.1. Variance reduction via variable reordering

Genz and Bretz (2009) suggested the following improvement of the SOV algorithm. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)$ be a permutation of the integers $1, \ldots, d$ and denote the corresponding permutation matrix P so that $P(1, \ldots, d)^{\mathrm{T}} = \boldsymbol{\pi}$. It is clear that for any $\boldsymbol{\pi}$ we have $l = \mathbb{P}(P\mathbf{l} \leqslant PA\mathbf{Z} \leqslant P\mathbf{u})$. Hence, to estimate l, we can input in the SOV algorithm 1 the permuted bounds and matrix: $\mathbf{l} \leftarrow P\mathbf{l}, \mathbf{u} \leftarrow P\mathbf{u}$ and $A \leftarrow PA$. This results in an unbiased estimator $\mathring{l}(\boldsymbol{\pi})$ whose variance will depend on $\boldsymbol{\pi}$—the order in which this high dimensional integration is carried out. Thus, we would like to choose the $\boldsymbol{\pi}^*$ among all possible permutations so that

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi}}{\operatorname{argmin}} \operatorname{var}\{\mathring{l}(\boldsymbol{\pi})\}.$$

This is an intractable combinatorial optimization problem whose objective function is not even available. Nevertheless, Genz and Bretz (2009) proposed a heuristic for finding an acceptable approximation to $\boldsymbol{\pi}^*$. We henceforth assume that this variable reordering heuristic is always applied as a preprocessing step to the SOV algorithm 1 so that the matrix A and the bounds $\mathbf{l}$ and $\mathbf{u}$ are already in permuted form. We shall revisit variable reordering in the numerical experiments in Section 5.

The main limitation of the estimator $\mathring{l}$ (with or without variable reordering) is that $\operatorname{var}(\mathring{l})$ is unknown and its estimate S^2 can be notoriously unreliable in the sense that the observed S^2 may be very small, whereas the true $\operatorname{var}(\mathring{l})$ is huge (Kroese *et al.*, 2011; Botev *et al.*, 2013). Such examples for which $\mathring{l}$ fails to deliver meaningful estimates of l will be given in the numerical Section 5.

2.2. Accept–reject simulation

The SOV approach that was described above suggests that we could simulate from $f(\mathbf{z})$ exactly by using $g(\mathbf{x})$ as an instrumental density in the accept–reject scheme in Table 2 (Kroese *et al.* (2011), chapter 3).

Of course, the accept–reject scheme will only be usable if the probability of acceptance $\mathbb{P}_{\mathbf{0}}\{cU \leqslant \phi(\mathbf{X}; \mathbf{0}, I)/g(\mathbf{X})\} = l/c$ is high and simulation from g is fast. Thus, this scheme presents two significant challenges which need resolution. The first is the computation of the constant c (or a very tight upper bound of it) in finite time. Locating the global maximum of the likelihood ratio $\phi(\mathbf{x}; \mathbf{0}, I)/g(\mathbf{x})$ may be an intractable problem—a local maximum will yield an incorrect sampling scheme. The second challenge is to select an instrumental g so that the acceptance probability is not prohibitively small (a ‘rare event’ probability). Unfortunately, the obvious choice

(4) resolves neither of these challenges (Hajivassiliou and McFadden, 1998). Other accept–reject schemes (Chopin, 2011), although excellent in one and two dimensions, ultimately have acceptance rates of the order $\mathscr{O}(2^{1-d})$ rendering them unusable for this type of problem with, say, $d=100$. We now address these issues concurrently in the next section.

3. Minimax tilting

Exponential tilting is a prominent technique in simulation (L'Ecuyer *et al.*, 2010; Kroese *et al.*, 2011). For a given light-tailed probability density $h(y)$ on $\mathbb{R}$, we can associate with h its exponentially tilted version $h_\mu(y)=\exp\{\mu y-K(\mu)\}h(y)$, where $K(\mu)=\ln[\mathbb{E}\{\exp(\mu X)\}]<\infty$, for some μ in an open set, is the cumulant-generating function. For example, the exponentially tilted version of $\phi(\mathbf{x};\mathbf{0},I)$ is $\exp\{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}-K(\boldsymbol{\mu})\}\phi(\mathbf{x};\mathbf{0},I)=\phi(\mathbf{x};\boldsymbol{\mu},I)$. Similarly, the tilted version of expression (4) yields

$$g_k(x_k;\mu_k|x_1,\ldots,x_{k-1})=\frac{\phi(x_k;\mu_k,1)\mathbb{I}(\tilde{l}_k\leqslant x_k\leqslant\tilde{u}_k)}{\Phi(\tilde{u}_k-\mu_k)-\Phi(\tilde{l}_k-\mu_k)}. \tag{5}$$

To simplify the notation in the subsequent analysis, let

$$\psi(\mathbf{x};\boldsymbol{\mu})\overset{\text{def}}{=}-\mathbf{x}^{\mathrm{T}}\boldsymbol{\mu}+\frac{\|\boldsymbol{\mu}\|^2}{2}+\sum_k\ln[\Phi\{\tilde{u}_k(x_1,\ldots,x_{k-1})-\mu_k\}-\Phi\{\tilde{l}_k(x_1,\ldots,x_{k-1})-\mu_k\}]. \tag{6}$$

Then, the tilted version of estimator (2) can be written as $\hat{l}=\exp\{\psi(\mathbf{X};\boldsymbol{\mu})\}$ with $\mathbf{X}\sim\mathbb{P}_{\boldsymbol{\mu}}$, where $\mathbb{P}_{\boldsymbol{\mu}}$ is the measure with PDF

$$g(\mathbf{x};\boldsymbol{\mu})\overset{\text{def}}{=}\prod_{k=1}^{d}g_k(x_k;\mu_k|x_1,\ldots,x_{k-1}).$$

It is now clear that the statistical properties of $\hat{l}$ depend on the tilting parameter $\boldsymbol{\mu}$. There is a large literature on the best way to select the tilting parameter $\boldsymbol{\mu}$; see L'Ecuyer *et al.* (2010) and the references therein. A recurrent theme in all works is the efficiency of the estimator $\hat{l}$ in a tail asymptotic regime where $l\downarrow 0$ is a rare event probability—precisely the setting that makes current accept–reject schemes inefficient. Thus, before we continue, we briefly recall the three widely used criteria for assessing efficiency in estimating tail probabilities.

The weakest type of efficiency and the most commonly encountered in the design of importance sampling schemes (Kroese *et al.*, 2011) is logarithmic efficiency. The estimator $\hat{l}$ is said to be *logarithmically or weakly efficient* if

$$\liminf_{l\downarrow 0}\frac{\ln\{\mathrm{var}(\hat{l})\}}{\ln(l^2)}\geqslant 1.$$

The second and stronger type of efficiency is the *bounded relative error*,

$$\limsup_{l\downarrow 0}\frac{\mathrm{var}(\hat{l})}{\hat{l}^2}\leqslant\text{constant}<\infty.$$

Finally, the best that we can hope for in an asymptotic regime is the highly desirable VRE property:

$$\limsup_{l\downarrow 0}\frac{\mathrm{var}(\hat{l})}{\hat{l}^2}=0.$$

An estimator is *strongly efficient* if it exhibits either bounded relative error or VRE. To achieve one of these efficiency criteria, most methods (L'Ecuyer *et al.*, 2010) rely on the derivation of an analytical asymptotic approximation to the relative error $\mathrm{var}(\hat{l})/l^2$, whose behaviour is then controlled by using the tilting parameter. The strongest type of efficiency VRE is uncommon for light-tailed probabilities and is typically achieved only within a state-dependent importance sampling framework (L'Ecuyer *et al.*, 2010).

Here we take a different tack: one that exploits features that are unique to the problem at hand and that will yield gains in efficiency in both an asymptotic and a non-asymptotic regime. A key result in this direction is the following lemma 1, whose proof is given in Appendix A.

Lemma 1 (minimax tilting). The optimization program

$$\inf_{\boldsymbol{\mu}} \sup_{\mathbf{x}\in\mathscr{C}} \psi(\mathbf{x};\boldsymbol{\mu})$$

is a saddle point problem with a unique solution given by the concave optimization program:

$$(\mathbf{x}^*,\boldsymbol{\mu}^*) = \underset{\mathbf{x},\boldsymbol{\mu}}{\arg\max}\ \psi(\mathbf{x};\boldsymbol{\mu}) \qquad \text{subject to } \frac{\partial\psi}{\partial\boldsymbol{\mu}} = \mathbf{0}, \quad \mathbf{x}\in\mathscr{C}. \tag{7}$$

Note that problem (7) minimizes with respect to $\boldsymbol{\mu}$ the worst case behaviour of the likelihood ratio, namely $\sup_{\mathbf{x}\in\mathscr{C}} \exp\{\psi(\mathbf{x};\boldsymbol{\mu})\}$. Lemma 1 states that we can both easily locate the global worst case behaviour of the likelihood ratio and simultaneously locate (in finite computing time) the global minimum with respect to $\boldsymbol{\mu}$. Before analysing the theoretical properties of minimax tilting, we first explain how to implement the minimax method in practice.

3.1. Practical implementation

How do we find the solution of problem (7) numerically? Without the constraint $\mathbf{x}\in\mathscr{C}$, the solution to problem (7) would be obtained by solving the non-linear system of equations $\nabla\psi(\mathbf{x};\boldsymbol{\mu}) = \mathbf{0}$, where the gradient is with respect to the vector $(\mathbf{x},\boldsymbol{\mu})$. To show why this is so, we introduce the following notation. Let $D = \mathrm{diag}(L)$, $\check{L} = D^{-1}L$ and

$$\Psi_j \overset{\text{def}}{=} \frac{\phi(\tilde{l}_j;\mu_j,1) - \phi(\tilde{u}_j;\mu_j,1)}{\mathbb{P}(\tilde{l}_j-\mu_j \leqslant Z \leqslant \tilde{u}_j-\mu_j)},$$

$$\Psi'_j \overset{\text{def}}{=} \frac{\partial\Psi_j}{\partial\mu_j} = \frac{(\tilde{l}_j-\mu_j)\,\phi(\tilde{l}_j;\mu_j,1) - (\tilde{u}_j-\mu_j)\,\phi(\tilde{u}_j;\mu_j,1)}{\mathbb{P}(\tilde{l}_j-\mu_j \leqslant Z \leqslant \tilde{u}_j-\mu_j)} - \Psi_j^2.$$

Then, the gradient equation $\nabla\psi(\mathbf{x};\boldsymbol{\mu}) = \mathbf{0}$ can be written as

$$\begin{aligned} \partial\psi/\partial\mathbf{x} &= -\boldsymbol{\mu} + (\check{L}^{\mathrm{T}} - I)\boldsymbol{\Psi} = \mathbf{0},\\ \partial\psi/\partial\boldsymbol{\mu} &= \boldsymbol{\mu} - \mathbf{x} + \boldsymbol{\Psi} = \mathbf{0}, \end{aligned} \tag{8}$$

and the Jacobian matrix has elements

$$\left.\begin{aligned} \partial^2\psi/\partial\boldsymbol{\mu}^2 &= I + \mathrm{diag}(\boldsymbol{\Psi}'),\\ \partial^2\psi/\partial\boldsymbol{\mu}\,\partial\mathbf{x} &= (\check{L} - I)\,\mathrm{diag}(\boldsymbol{\Psi}') - I,\\ \partial^2\psi/\partial\mathbf{x}^2 &= (\check{L} - I)^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\Psi}')(\check{L} - I). \end{aligned}\right\} \tag{9}$$

The Karush–Kuhn–Tucker equations give the necessary and sufficient condition for the global solution $(\mathbf{x}^*,\boldsymbol{\mu}^*)$ of problem (7):

Table 3. Algorithm 3 (computation of optimal pair $(\mathbf{x}^*, \boldsymbol{\mu}^*)$)

```
Use Powell's (1970) dogleg method on equations (8) with Jacobian (9) to find $(\check{\mathbf{x}}, \check{\boldsymbol{\mu}})$:
if $(\check{\mathbf{x}}, (\check{\boldsymbol{\mu}})) \in \mathscr{C}$ then
    $(\mathbf{x}^*, \boldsymbol{\mu}^*) \leftarrow (\check{\mathbf{x}}, \check{\boldsymbol{\mu}})$;
else
    use a convex solver to find $(\mathbf{x}^*, \boldsymbol{\mu}^*)$, where $(\check{\mathbf{x}}, \check{\boldsymbol{\mu}})$ is the initial guess;
return $(\mathbf{x}^*, \boldsymbol{\mu}^*)$
```

$$\left.\begin{array}{lll} \partial\psi/\partial\boldsymbol{\mu}=\mathbf{0}, & \partial\psi/\partial\mathbf{x}-\check{L}^{\mathrm{T}}\boldsymbol{\eta}_1+\check{L}^{\mathrm{T}}\boldsymbol{\eta}_2=\mathbf{0}, & \\ \boldsymbol{\eta}_1\geqslant\mathbf{0}, & L\mathbf{x}-\mathbf{u}\leqslant\mathbf{0}, & \boldsymbol{\eta}_1^{\mathrm{T}}(L\mathbf{x}-\mathbf{u})=\mathbf{0}, \\ \boldsymbol{\eta}_2\geqslant\mathbf{0}, & -L\mathbf{x}+\mathbf{l}\leqslant\mathbf{0}, & \boldsymbol{\eta}_2^{\mathrm{T}}(L\mathbf{x}-\mathbf{l})=\mathbf{0}, \end{array}\right\} \qquad (10)$$

where $\boldsymbol{\eta}_1$and $\boldsymbol{\eta}_2$ are Lagrange multipliers.

Suppose that we find the unique solution of the non-linear system (8) by using, for example, a trust region dogleg method (Powell, 1970). If we denote the solution to equations (8) by $(\check{\mathbf{x}}, \check{\boldsymbol{\mu}})$, then the Karush–Kuhn–Tucker equations imply that $(\check{\mathbf{x}}, \check{\boldsymbol{\mu}}) = (\mathbf{x}^*, \boldsymbol{\mu}^*)$ if and only if $(\check{\mathbf{x}}, \check{\boldsymbol{\mu}}) \in \mathscr{C}$ or equivalently $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2 = \mathbf{0}$. If, however, the solution $(\check{\mathbf{x}}, \check{\boldsymbol{\mu}})$ to equations (8) does not lie in $\mathscr{C}$, then $(\check{\mathbf{x}}; \check{\boldsymbol{\mu}})$ will be suboptimal and, to compute $(\mathbf{x}^*; \boldsymbol{\mu}^*)$, we must use a constrained convex optimization solver. This observation then leads to the procedure in Table 3.

Numerical experience suggests that almost always $(\check{\mathbf{x}}, \check{\boldsymbol{\mu}})$ happens to lie in $\mathscr{C}$ and there is no need to do any additional computation beyond Powell's (1970) trust region method.

4. Theoretical properties of minimax tilting

There are several reasons why the minimax program (7) is an excellent way of selecting the tilting parameter. The first shows that, unlike its competitors, the estimator proposed,

$$\hat{l} = \exp\{\psi(\mathbf{X}; \boldsymbol{\mu}^*)\}, \qquad \mathbf{X} \sim \mathbb{P}_{\boldsymbol{\mu}^*}, \qquad (11)$$

achieves the best possible efficiency in a tail asymptotic regime.

Let $\Sigma = AA^{\mathrm{T}}$ be a full rank covariance matrix. Consider the tail probability $l(\gamma) = \mathbb{P}(\mathbf{X} \geqslant \gamma\mathbf{l})$, where $\mathbf{X} \sim N(\mathbf{0}, \Sigma)$ and $\gamma > 0$ and $\mathbf{l} > \mathbf{0}$. We show that estimator (11) exhibits strong efficiency in estimating $l(\gamma)$ as $\gamma \uparrow \infty$. For this, we first introduce the following simplifying notation.

Similarly to the variable reordering in Section 2.1, suppose that P is a permutation matrix which maps the vector $(1, \ldots, d)^{\mathrm{T}}$ into the permutation $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)^{\mathrm{T}}$, i.e. $P(1, \ldots, d)^{\mathrm{T}} = \boldsymbol{\pi}$. Let L be the lower triangular factor of $P\Sigma P^{\mathrm{T}} = LL^{\mathrm{T}}$ and $\mathbf{p} = P\mathbf{l}$. It is clear that

$$l(\gamma) = \mathbb{P}(P\mathbf{X} \geqslant \gamma P\mathbf{l}) = \mathbb{P}(L\mathbf{Z} \geqslant \gamma\mathbf{p})$$

for any permutation $\boldsymbol{\pi}$. For the time being, we leave $\boldsymbol{\pi}$ unspecified, because, unlike in Section 2.1, here we do not use $\boldsymbol{\pi}$ to minimize the variance of the estimator, but to simplify the notation in our efficiency analysis.

Define the convex quadratic programming problem

$$\min_{\mathbf{x}} \tfrac{1}{2} \|\mathbf{x}\|^2 \qquad \text{subject to } L\mathbf{x} \geqslant \gamma\mathbf{p}. \qquad (12)$$

The Karush–Kuhn–Tucker equations, which are a necessary and sufficient condition to find the solution of problem (12), are given by

$$\left.\begin{array}{c} \mathbf{x}-L^{\mathrm{T}}\boldsymbol{\lambda}=\mathbf{0}, \\ \boldsymbol{\lambda}\geqslant\mathbf{0}, \\ \gamma\mathbf{p}-L\mathbf{x}\leqslant\mathbf{0}, \\ \boldsymbol{\lambda}^{\mathrm{T}}(\gamma\mathbf{p}-L\mathbf{x})=0, \end{array}\right\} \tag{13}$$

where $\boldsymbol{\lambda}\in\mathbb{R}^d$ is a Lagrange multiplier vector. Suppose that the number of active constraints in problem (12) is d_1 and the number of inactive constraints is d_2, where $d_1+d_2=d$. Since $L\mathbf{x}\geqslant\gamma\mathbf{p}>\mathbf{0}$, the number of active constraints $d_1\geqslant 1$, because otherwise $\mathbf{x}=\mathbf{0}$ and $L\mathbf{x}=\mathbf{0}$, reaching a contradiction.

Given the partition $\boldsymbol{\lambda}=(\boldsymbol{\lambda}_1^{\mathrm{T}},\boldsymbol{\lambda}_2^{\mathrm{T}})^{\mathrm{T}}$ with $\dim(\boldsymbol{\lambda}_1)=d_1$ and $\dim(\boldsymbol{\lambda}_2)=d_2$, we now choose $\boldsymbol{\pi}$ such that all the active constraints in expression (13) correspond to $\boldsymbol{\lambda}_1>\mathbf{0}$ and all the inactive constraints to $\boldsymbol{\lambda}_2=\mathbf{0}$. Similarly, we define a partitioning for $\mathbf{x}$ and $\mathbf{p}$, and the lower triangular

$$L=\begin{pmatrix} L_{11} & O \\ L_{21} & L_{22} \end{pmatrix}.$$

The only reason for introducing the above variable reordering via the permutation matrix P and insisting that all active constraints of problem (12) are collected in the upper part of vector $\boldsymbol{\lambda}$ is notational convenience and simplicity. At the cost of some generality, this preliminary variable reordering allows us to state and prove the efficiency result in the following theorem 1 in its simplest and neatest form.

Theorem 1 (strong efficiency of minimax estimator). Consider the estimation of the probability

$$l(\gamma)=\mathbb{P}(\mathbf{X}\geqslant\gamma\mathbf{l})=\mathbb{P}(L\mathbf{Z}\geqslant\gamma\mathbf{p})$$

where $\mathbf{X}\sim N(\mathbf{0},\Sigma)$ and $\mathbf{Z}\sim N(\mathbf{0},I)$, and $LL^{\mathrm{T}}=P\Sigma P^{\mathrm{T}},\mathbf{p}=P\mathbf{l}>\mathbf{0}$ are the permuted versions of Σ and $\mathbf{l}$ ensuring that the Lagrange multiplier vector $\boldsymbol{\lambda}$ in expression (13) satisfies $\boldsymbol{\lambda}_1>\mathbf{0}$ and $\boldsymbol{\lambda}_2=\mathbf{0}$. Define

$$\mathbf{q}\overset{\text{def}}{=}L_{21}L_{11}^{-1}\mathbf{p}_1-\mathbf{p}_2$$

and let $\mathscr{J}$ be the set of indices for which the components of the vector $\mathbf{q}$ are 0, i,e.

$$\mathscr{J}\overset{\text{def}}{=}\{j:q_j=0,j=1,\ldots,d_2\}. \tag{14}$$

If $\mathscr{J}=\emptyset$, then the minimax estimator (11) is a VRE estimator:

$$\limsup_{\gamma\uparrow\infty}\frac{\mathrm{var}_{\boldsymbol{\mu}^*}\{\hat{l}(\gamma)\}}{l^2(\gamma)}=0.$$

Alternatively, if $\mathscr{J}\neq\emptyset$, then $\hat{l}$ is a bounded relative error estimator:

$$\limsup_{\gamma\uparrow\infty}\frac{\mathrm{var}_{\boldsymbol{\mu}^*}\{\hat{l}(\gamma)\}}{l^2(\gamma)}<\text{constant}<\infty.$$

Theorem 1 suggests that, unless the covariance matrix Σ has a very special structure, the estimator enjoys VRE. This raises the question: is there a simple setting that guarantees VRE for any full rank covariance matrix under any preliminary variable reordering?

The next result shows that, when $\mathbf{l}$ can be represented as a weighted linear combination of the columns of the covariance matrix $\Sigma=AA^{\mathrm{T}}$, then we always have VRE.

Theorem 2 (minimax VRE). Consider the estimation of the tail probability $l(\gamma)=\mathbb{P}(\gamma\mathbf{l}\leqslant A\mathbf{Z}\leqslant\infty)$, where $\mathbf{l}=\Sigma\mathbf{l}^*$ for some positive weight $\mathbf{l}^*>\mathbf{0}$. Then, the minimax estimator (11) is a VRE estimator.

In contrast, under the additional assumption $L^{\mathrm{T}}\mathbf{l}^*>\mathbf{0}$ (strong positive covariance), where L is the lower triangular factor of $\Sigma=LL^{\mathrm{T}}$, the SOV estimator $\mathring{l}$ is a bounded relative error estimator; otherwise, it is a divergent relative error estimator:

$$\frac{\mathrm{var}_{\mathbf{0}}[\exp\{\psi(\mathbf{X};\mathbf{0})\}]}{l^2(\gamma)}\simeq\begin{cases}\mathcal{O}(1) & \text{if } L^{\mathrm{T}}\mathbf{l}^*>\mathbf{0},\\ \exp[\mathcal{O}(\gamma^2)+\mathcal{O}\{\ln(\gamma)\}+\mathcal{O}(1)] & \text{otherwise.}\end{cases}$$

(The symbols $f(x)\simeq g(x)$, $f(x)=\mathcal{O}\{g(x)\}$ and $f(x)=o\{g(x)\}$, as $x\uparrow\infty$ and $g(x)\neq 0$, stand for $\lim_{x\uparrow\infty}f(x)/g(x)=1$, $\limsup_{x\uparrow\infty}|f(x)/g(x)|<\infty$ and $\lim_{x\uparrow\infty}f(x)/g(x)=0$ respectively.)

Note that the permutation matrix P plays no role in the statement of theorem 2 (we can assume that $P=I$), and that we do not assume $\mathbf{l}>\mathbf{0}$, but only that $\mathbf{l}=\Sigma\mathbf{l}^*$ for some $\mathbf{l}^*>\mathbf{0}$.

In light of theorems 1 and 2, for the obverse problem of simulation from the truncated multivariate normal distribution we obtain the following result.

Corollary 1 (asymptotically efficient simulation). Suppose that the instrumental density in the accept–reject algorithm 2 for simulation from

$$f(\mathbf{z})\propto\phi(\mathbf{z};\mathbf{0},I)\,\mathbb{I}(A\mathbf{z}\geqslant\gamma\mathbf{l})$$

is given by $g(\mathbf{x};\boldsymbol{\mu}^*)$. Suppose further that either $\mathbf{l}>\mathbf{0}$ and the corresponding estimator (11) enjoys VRE, or $\mathbf{l}=\Sigma\mathbf{l}^*$ for some $\mathbf{l}^*>\mathbf{0}$. Then, the measure $\mathbb{P}_{\boldsymbol{\mu}^*}$ becomes indistinguishable from the target $\mathbb{P}$:

$$\sup_{\mathscr{A}}|\mathbb{P}(\mathbf{Z}\in\mathscr{A})-\mathbb{P}_{\boldsymbol{\mu}^*}(\mathbf{Z}\in\mathscr{A})|\rightarrow 0,\qquad \gamma\uparrow\infty.$$

A second reason that recommends our choice of tilting parameter is that $\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}$ is a non-trivial deterministic upper bound to l, i.e. $l\leqslant\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}$.

As a result, unlike many existing estimators (Vijverberg, 1997; Genz, 1992), we can construct an exact (albeit conservative) confidence interval for l as follows. Let $\varepsilon>0$ be the desired width of the $1-\alpha$ confidence interval and $l_{\mathrm{L}}\leqslant l$ be a lower bound to l. Then, by Hoeffding's inequality for $\bar{l}=(\hat{l}_1+\ldots+\hat{l}_n)/n$ with

$$n(\varepsilon)=\lceil -\ln(\alpha/2)[\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}-l_{\mathrm{L}}]^2/(2\varepsilon^2)\rceil,\tag{15}$$

we obtain $\mathbb{P}_{\boldsymbol{\mu}^*}(\bar{l}-\varepsilon\leqslant l\leqslant\bar{l}+\varepsilon)\geqslant 1-\alpha$. As is widely known (Kroese *et al.*, 2011), the main weakness of any importance sampling estimator $\bar{l}$ of l is the risk of severe underestimation of l. Thus, plugging in $\bar{l}$ (or, even more conservatively, plugging in 0) in place of l_{L} in the formula for n above will yield a robust confidence interval $\bar{l}\pm\varepsilon$. For practitioners who are not satisfied with such a heuristic approach, we provide the following deterministic lower bound to l.

Lemma 2 (cross-entropy lower bound). Define the product measure $\mathbb{P}$ with PDF

$$\phi(\mathbf{x})\propto\phi\{\mathbf{x};\boldsymbol{\nu},\mathrm{diag}^2(\boldsymbol{\sigma})\}\,\mathbb{I}(\mathbf{l}\leqslant\mathbf{x}\leqslant\mathbf{u}),$$

where $\boldsymbol{\nu}$ and $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_d)^{\mathrm{T}}$ are location and scale parameters respectively. Define

$$l_{\mathrm{L}}=\sup_{\boldsymbol{\nu},\boldsymbol{\sigma}}\frac{\exp(-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}\underline{\mathrm{var}}(\mathbf{X})\}-\frac{1}{2}\underline{\mathbb{E}}[\mathbf{X}]^{\mathrm{T}}\Sigma^{-1}\underline{\mathbb{E}}[\mathbf{X}]-\underline{\mathbb{E}}[\ln\{\underline{\phi}(\mathbf{X})\}]}{(2\pi)^{d/2}|\det(A)|},$$

where $\Sigma=AA^{\mathrm{T}}$. Then, $l_{\mathrm{L}}\leqslant l$ is a variational lower bound to l. In addition, under the conditions of theorem 2, namely $(\mathbf{l},\mathbf{u})=(\gamma\Sigma\mathbf{l}^*,\infty)$, we have that $l_{\mathrm{L}}\uparrow l(\gamma)$ and

$$\sup_{\mathscr{A}} |\mathbb{P}(\mathbf{Z} \in \mathscr{A}) - \underline{\mathbb{P}}(A^{-1}\mathbf{Z} \in \mathscr{A})| \downarrow 0, \qquad \gamma \uparrow \infty. \tag{16}$$

Since simulation from $\underline{\mathbb{P}}$ is straightforward, one may be tempted to consider using $\underline{\mathbb{P}}$ as an alternative importance measure to $\mathbb{P}_{\mu^*}$. Unfortunately, despite the similarity of the results in theorem 2 and lemma 2, the PDF $\underline{\phi}$ is not amenable to an accept–reject scheme for exact sampling from f and as an importance sampling measure it does not yield VRE. Thus, the sole use of lemma 1 is for constructing an exact confidence interval and lower bound to l in the tails of the normal distribution.

Under the conditions of theorem 2, the minimax estimator enjoys the bounded normal approximation property (Tuffin, 1999). i.e. if $\bar{l}$ and S^2 are the mean and sample variance of the IID $\hat{l}_1, \ldots, \hat{l}_n$, and $F_n(x)$ is the empirical cumulative density function of $T_n = \sqrt{n}(\bar{l} - l)/S$, then we have the Berry–Esséen bound, uniformly in γ:

$$\sup_{x \in \mathbb{R}, \gamma > 0} |F_n(x) - \Phi(x)| \leqslant \text{constant}/\sqrt{n}.$$

This Berry–Esséen bound implies that the coverage error of the approximate $(1-\alpha)$-level confidence interval $\bar{l} \pm z_{1-\alpha/2} S/\sqrt{n}$ remains of the order $\mathcal{O}(n^{-1/2})$, even as $l \downarrow 0$. Thus, if a lower bound l_{L} is not easily available, one can still rely on the confidence interval that is derived from the central limit theorem.

Finally, in addition to the strong efficiency properties of the estimator, another reason that recommends the minimax estimator is that it permits us to tackle intractable simulation and estimation problems as illustrated in the next section.

5. Numerical examples and applications

We begin by considering some test cases that have been used throughout the literature (Fernández *et al.*, 2007; Craig, 2008; Miwa *et al.*, 2003). We are interested in both the efficient simulation of the Gaussian vector $\mathbf{X} = A\mathbf{Z} \sim N(\mathbf{0}, \Sigma)$ conditionally on $\mathbf{X} \in \mathscr{A}$, and the estimation of l in expression (1).

In all examples we compare the SOV estimator of Genz with the proposed minimax exponentially tilted (MET) estimator. We note that initially we considered a comparison with other estimation schemes such as the radially symmetric approach of Nomura (2014) and the specialized orthant probability algorithm of Miwa *et al.* (2003), Craig (2008) and Nomura (2016). Unfortunately, unless a special auto-regressive covariance structure is present, these methods are hardly competitive in anything but very few dimensions. For example, the orthant algorithm of Miwa *et al.* (2003) has complexity $\mathcal{O}(d!n)$, which becomes too costly for $d > 10$. For this reason, we give a comparison only with the broadly applicable SOV scheme, which is widely recognized as the current state of the art method.

Since both the SOV and the MET estimators are smooth, one can seek further gains in efficiency by using a randomized quasi-Monte-Carlo algorithm. The idea behind the quasi-Monte-Carlo method is to reduce the error of the estimator by using *quasi-random* or *low discrepancy* sequences of numbers, instead of the traditional (pseudo)random sequences. Typically the error of a sample average estimator decays at the rate of $\mathcal{O}(n^{-1/2})$ when using random numbers, and at the rate of $\mathcal{O}\{\ln(n)^d/n\}$ when using pseudorandom numbers; see Gerber and Chopin (2015) for an up-to-date discussion.

For both the SOV and the MET estimator we use the n-point Richtmyer quasi-random sequence with randomization, as recommended by Genz and Bretz (2009). The randomization allows us to estimate the variability of the estimator in the standard Monte Carlo manner. The details are summarized in Table 4.

Table 4. Algorithm 4 (randomized quasi-Monte-Carlo algorithm (Genz and Bretz, 2009))

```
*Require* dimension $d$ and sample size $n$:
    $d' \leftarrow \lceil 5d\ \ln(d+1)/4 \rceil$, $n' \leftarrow \lceil n/12 \rceil$;
    let $p_1, \ldots, p_{d'}$ be the first $d'$ prime numbers,
    $\mathbf{q}_i \leftarrow \surd p_i (1, \ldots, n')^{\mathrm{T}}$ for $i = 1, \ldots, d'$
    *for* $k = 1, \ldots, 12$ *do*
        *for* $i = 1, \ldots, d-1$ do
            *let* $U \sim U(0,1)$, independently
            $\mathbf{s}_i \leftarrow |2\{(\mathbf{q}_i + U) \bmod 1\} - 1|$
        $qms \leftarrow (\mathbf{s}_1, \ldots, \mathbf{s}_{d-1})$;
        use the sequence $qms$ to compute an $n'$-point sample average estimator $\hat{l}_k$;
*return* $\bar{l} \leftarrow (1/12)\Sigma_k \hat{l}_k$ with estimated relative error $(1/12)\surd\{\Sigma_k(\hat{l}_k - \bar{l})^2\}/\bar{l}$
```

Table 5. Estimates of l for various values of d using $n = 10^4$ replications

d	l_{L}	*SOV estimator*	*MET estimator*	$\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}$	*Worst error (%)*	*Acceptance proportion*
2	0.0148955	0.0148963 (4×10^{-4}%)	0.01489 (4×10^{-5}%)	0.0149	2×10^{-4}	0.99
3	0.0010771	0.0010772 (3×10^{-3}%)	0.001077 (3×10^{-4}%)	0.00108	6×10^{-3}	0.99
5	2.4505×10^{-6}	2.4508×10^{-6} (0.08%)	2.451×10^{-6} (0.002%)	2.48×10^{-6}	0.012	0.98
10	8.5483×10^{-15}	8.4591×10^{-15} (0.8%)	8.556×10^{-15} (0.01%)	2.1046×10^{-14}	0.03	0.97
15	1.3717×10^{-25}	1.366×10^{-25} (11%)	1.375×10^{-25} (0.01%)	1.43×10^{-25}	0.04	0.95
20	1.7736×10^{-38}	1.65×10^{-38} (37%)	1.7796×10^{-38} (0.03%)	1.869×10^{-38}	0.05	0.95
25	2.674×10^{-53}	2.371×10^{-48} (33%)	2.6847×10^{-53} (0.02%)	2.83×10^{-53}	0.05	0.94
30	6.09×10^{-70}	—	6.11×10^{-70} (0.03%)	6.46×10^{-70}	0.05	0.94
40	2.17×10^{-108}	—	2.18×10^{-108} (0.05%)	2.30×10^{-108}	0.06	0.94
50	2.1310×10^{-153}	—	2.1364×10^{-153} (0.06%)	2.24×10^{-153}	0.05	0.95

Since there is no need to integrate the x_dth component, the loop over i in algorithm 4 goes up to $d-1$.

5.1. Structured covariance matrices

At this junction we assume that the matrix A (or equivalently Σ) and the bounds $\mathbf{l}$ and $\mathbf{u}$ have already been permuted according to the variable reordering heuristic that was discussed in Section 2.1. Thus, the ordering of the variables during the integration will be the same for both estimators and will not matter in the comparison.

5.1.1. *Example I (Fernández et al., 2007)*

Consider $\mathscr{A} = [\frac{1}{2}, 1]^d$ with a covariance matrix

$$\Sigma^{-1} = \tfrac{1}{2} I + \tfrac{1}{2} \mathbf{1}\mathbf{1}^{\mathrm{T}}.$$

The third and fourth columns in Table 5 show the estimates of l for various values of d. The figure in parentheses give the estimated relative error as percentages.

The second column shows the lower bound that was discussed in lemma 2 and the fifth column shows the deterministic upper bound. These two bounds can then be used to compute the exact

Table 6. Estimates of l for various values of d using $n=10^4$ replications

d	l_{L}	*SOV estimator*	*MET estimator*	$\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}$	*Worst error (%)*	*Acceptance proportion*
2	0.09114	0.09121 (6×10^{-4}%)	0.09121 (2×10^{-4}%)	0.09205	0.009	0.99
3	0.02303	0.02307 (0.001%)	0.02307 (4×10^{-4}%)	0.0234	0.01	0.98
10	1.338×10^{-6}	1.3493×10^{-6} (0.03%)	1.3490×10^{-6} (0.003%)	1.454×10^{-6}	0.07	0.92
20	1.080×10^{-12}	1.0982×10^{-12} (0.23%)	1.0989×10^{-12} (0.004%)	1.289×10^{-12}	0.17	0.85
25	9.770×10^{-16}	1.00×10^{-15} (0.28%)	9.9808×10^{-16} (0.02%)	1.222×10^{-15}	0.2	0.81
50	5.925×10^{-31}	6.137×10^{-31} (0.7%)	6.188×10^{-31} (0.05%)	9.368×10^{-31}	0.5	0.66
80	3.252×10^{-49}	3.477×10^{-49} (1.8%)	3.479×10^{-49} (0.1%)	6.812×10^{-49}	1.0	0.50
100	2.18×10^{-61}	2.351×10^{-61} (3%)	2.384×10^{-61} (0.2%)	5.50×10^{-61}	1.3	0.43
120	1.462×10^{-73}	1.641×10^{-73} (5.6%)	1.622×10^{-73} (0.3%)	4.45×10^{-73}	1.7	0.36
150	8.026×10^{-92}	9.751×10^{-92} (6.3%)	9.142×10^{-92} (0.18%)	3.23×10^{-91}	2.5	0.28
200	2.954×10^{-122}	3.581×10^{-122} (11%)	3.525×10^{-122} (0.5%)	1.905×10^{-121}	4.4	0.18
250	1.087×10^{-152}	1.359×10^{-152} (15%)	1.357×10^{-152} (0.6%)	1.120×10^{-151}	7.2	0.12

confidence interval (mentioned in the previous section) whenever we allow n to vary freely. Here, since n is fixed and the error is allowed to vary, we instead display the upper bound to the relative error (given in the sixth column under the 'worst error' heading)

$$\frac{\sqrt{\mathrm{var}(\bar{l})}}{l} \leqslant \frac{\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}/l_{\mathrm{L}}-1}{\sqrt{n}}.$$

Finally, the seventh column ('acceptance proportion') gives the acceptance rate of algorithm 2 when using the instrumental density $g(\cdot;\boldsymbol{\mu}^*)$ with enveloping constant $c=\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}$.

What makes the MET approach better than other methods? First, the acceptance rate in the last column remains high even for $d=50$. In contrast, the acceptance rate from naive acceptance–rejection with instrumental PDF $\phi(\mathbf{0},\Sigma)$ is a rare event probability of approximately 2.13×10^{-153}. Note again that the existing accept–reject scheme of Chopin (2011) is an excellent algorithm designed for extremely fast simulation in one or two dimensions (in quite general settings) and is not suitable here.

Second, the performance of both the SOV and the MET estimators gradually deteriorates with increasing d. However, the SOV estimator has larger relative error, does not give meaningful results for $d>25$ and has no theoretical quantification of its performance. In contrast, the MET estimator is guaranteed to have better relative error than the estimator that is given in the sixth column (worst error).

Finally, in further numerical experiments (which are not displayed here) we observed that the width ε of the *exact* confidence interval $\bar{l}\pm\varepsilon$ with $\alpha=0.05$, based on the Hoeffding bound (15), was of the same order of magnitude as the width of the *approximate* confidence interval $\bar{l}\pm z_{1-\alpha/2}S/\sqrt{n(\varepsilon)}$.

5.1.2. Example II (Fernández et al., 2007)

Consider the hypercube $\mathscr{A}=[0,1]^d$ and the isotopic covariance with elements

$$(\Sigma^{-1})_{i,j}=\frac{1}{2^{|i-j|}}\mathbb{1}\left(|i-j|\leqslant\frac{d}{2}\right).$$

Observe how rapidly the probabilities in Table 6 become very small. Why should we be interested in estimating small 'rare event' probabilities? The simple answer is that all probabilities become eventually rare event probabilities as the dimensions grow increasingly larger, making naive accept–reject simulation infeasible. These small probabilities sometimes present not only theoretical challenges (rare event estimation) but also practical challenges like representation in finite precision arithmetic and numerical underflow. For instance, in using the SOV estimator Grün and Hornik (2012) noted that

> 'Numerical problems arise for very small probabilities, e.g. for observations from different components. To avoid these problems observations with a small posterior probability (smaller than or equal to 10^{-6}) are omitted in the M-step of this component.'

The MET estimator is not immune to numerical underflow and loss of precision during computation but, consistent with theorems 1 and 2, it is typically much more robust than the SOV estimator in estimating small probabilities.

5.2. Random-correlation matrices

One can argue that the covariance matrices that we have considered so far are too structured and hence not representative of a 'typical' covariance matrix. Thus, for simulation and testing Miwa *et al.* (2003) and Craig (2008) found it desirable to use random-correlation matrices. In the subsequent examples we use the method of Davies and Highman (2000) to simulate random-test correlation matrices whose eigenvalues are uniformly distributed over the simplex $\{\mathbf{x} : x_1 + \ldots + x_d = d\}$.

Table 7. Five-number summary for relative error based on 100 independent replications

	Minimum (%)	*1st quartile (%)*	*Median (%)*	*3rd quartile (%)*	*Maximum (%)*
MET estimator	0.07	0.12	0.17	0.20	0.44
SOV estimator	0.27	0.63	1.00	1.68	9.14
Acceptance rate	1.2	3.9	5.5	7.3	12

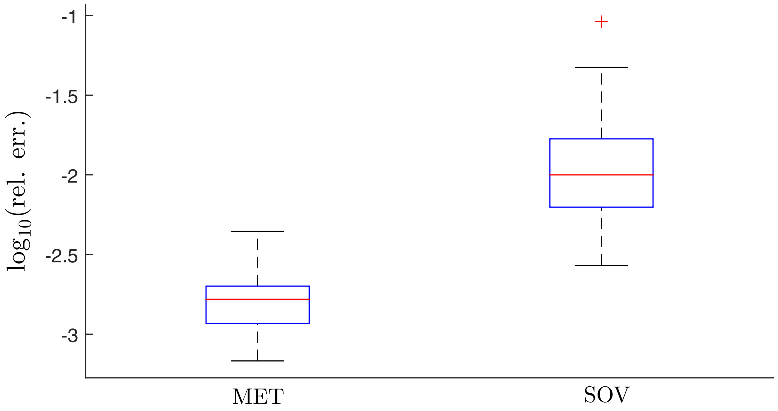

Fig. 1. Boxplots of relative errors from the outcomes of the 100 independent replications on a logarithmic scale in example III

Table 8. Relative errors of SOV and MET estimators over 100 random correlation cases

	Minimum (%)	*1st quartile (%)*	*Median (%)*	*3rd quartile (%)*	*Maximum (%)*
MET	0.020	0.044	0.077	0.12	0.44
SOV	4.3	15	26	48	99
Acceptance rate	1.5	10	18	26	43

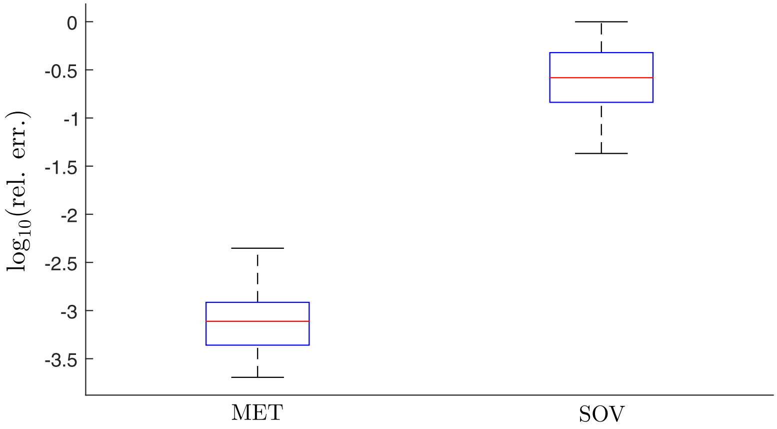

Fig. 2. Boxplots of relative errors from the outcomes of the 100 replications on a logarithmic scale in example IV

5.2.1. *Example III*

A natural question is whether the MET estimator would still be preferable when integrating over a 'non-tail' region such as $\mathscr{A}=[-\frac{1}{2},\infty]^{100}$. Table 7 and Fig. 1 summarize the output of running the algorithms on 100 independently simulated random-correlation matrices. Both the SOV and the MET estimators used $n=10^5$ quasi-Monte-Carlo points. The 'acceptance rate' row displays the five-number summary of the estimated acceptance probability of algorithm 2.

So far we have said little about the cost of computing the optimal pair $(\mathbf{x}^*;\boldsymbol{\mu}^*)$, and the measures of efficiency that we have considered do not account for the computational cost of the estimators. The reason for this is that, in the examples that we investigated, the computing time that is required to find the pair $(\mathbf{x}^*;\boldsymbol{\mu}^*)$ is insignificant compared with the time that it takes to evaluate $n>10^5$ replications of $\hat{l}$ or $\mathring{l}$.

In the current example, the numerical experiments suggest that the MET estimator is roughly 20% more costly than the SOV estimator. If we adjust the results in Table 1 to account for this time difference, then the relative error in the SOV row would be reduced by a factor of at most 1.2. This adjustment will thus give a reduction in the typical (median) relative error from 1.0 to $1/1.2\approx 0.83\%$, which is hardly significant.

5.2.2. *Example IV*

Finally, we wish to know whether the strong efficiency that was described in theorem 1 may benefit the MET estimator as we move further into the tails of the distribution. Choose the 'tail-like' $\mathscr{A} = [1, \infty]^{100}$ and use $n = 10^5$. Table 8 and Fig. 2 summarize the results of 100 replications.

As seen from the results, in this particular example the variance of the MET estimator is typically less than 10^5 the variance of the SOV estimator.

5.3. Computational limitations in high dimensions

It is important to emphasize the limitations of the minimax tilting approach. Like all other methods, including Markov chain Monte Carlo methods, it is not a panacea against the curse of dimensionality. The acceptance probability of algorithm 2 ultimately becomes a rare event probability as the dimensions keep increasing, because the bounded or vanishing relative error properties of $\hat{l}$ do not hold in the asymptotic regime $d \uparrow \infty$.

Numerical experiments suggest that the method generally works reliably for $d \leqslant 100$. The approach may sometimes be effective in higher dimensions provided that l does not decay too fast in d. In this regard, Miwa *et al.* (2003) and Craig (2008) studied the orthant probability $l = \mathbb{P}(\mathbf{X} \in [0, \infty]^d)$ with the positive correlation structure

$$\Sigma = \tfrac{1}{2} I + \tfrac{1}{2} \mathbf{1}\mathbf{1}^{\mathrm{T}}.$$

This is a rare case for which the exact value of the probability is known, namely $l = 1/(d+1)$, and decays very slowly to 0 as $d \uparrow \infty$. For this reason, we use it to illustrate the behaviour of the SOV and MET estimators for very large d.

Fig. 3 and Table 9 show the output of a numerical experiment with $n = 10^5$ for various values of d. Fig. 3 gives the computational cost in seconds. Both the SOV and the MET estimators have a cost of $\mathcal{O}(d^3)$—hence the excellent agreement with the least squares cubic polynomials fitted to the empirical central processor unit data. Table 9 displays the relative error for both methods. In this example, we apply the variable reordering heuristic to the SOV estimator only, illustrating that the heuristic is not always necessary to achieve satisfactory performance with the MET estimator.

This example confirms the result in theorem 2 that the SOV estimator works better in settings with strongly positive correlation structure (but poorly with negative correlation). Further, the results suggest that the MET estimator is also aided by the presence of positive correlation.

5.4. Exact simulation of probit posterior

A popular generalized linear model (Koop *et al.*, 2007) for binary responses $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$ with explanatory variables $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{ik})^{\mathrm{T}}$, $i = 1, \ldots, m$, is the probit Bayesian model:

(a) prior, $p(\boldsymbol{\beta}) \propto \exp\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}} V^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}$ with $\boldsymbol{\beta} \in \mathbb{R}^k$ and for simplicity $\boldsymbol{\beta}_0 = \mathbf{0}$;
(b) likelihood, $p(\mathbf{y}|\boldsymbol{\beta}) \propto \exp(\Sigma_{i=1}^m \ln[\Phi\{(2y_i - 1)\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\}])$.

The challenge is to simulate from the posterior $p(\boldsymbol{\beta}|\mathbf{y})$. We can use latent variables (Albert and Chib, 1993) to represent the posterior as the marginal of a truncated multivariate normal distribution. Let $\boldsymbol{\lambda} \sim N(0, I_m)$ be latent variables and define the design matrix $\tilde{X} = \mathrm{diag}(2\mathbf{y} - \mathbf{1})X$. Then, the marginal $f(\boldsymbol{\beta})$ of the joint PDF

$$f(\boldsymbol{\beta}, \boldsymbol{\lambda}) \propto \exp(-\tfrac{1}{2} \| V^{-1/2} \boldsymbol{\beta} \|^2 - \tfrac{1}{2} \| \boldsymbol{\lambda} \|^2)\, \mathbb{I}(\tilde{X}\boldsymbol{\beta} - \boldsymbol{\lambda} \geqslant \mathbf{0})$$

equals the desired posterior $p(\boldsymbol{\beta}|\mathbf{y})$. We can thus apply our accept–reject scheme, because the joint $f(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is of the desired truncated multivariate form (1) with $d = k + m$ and

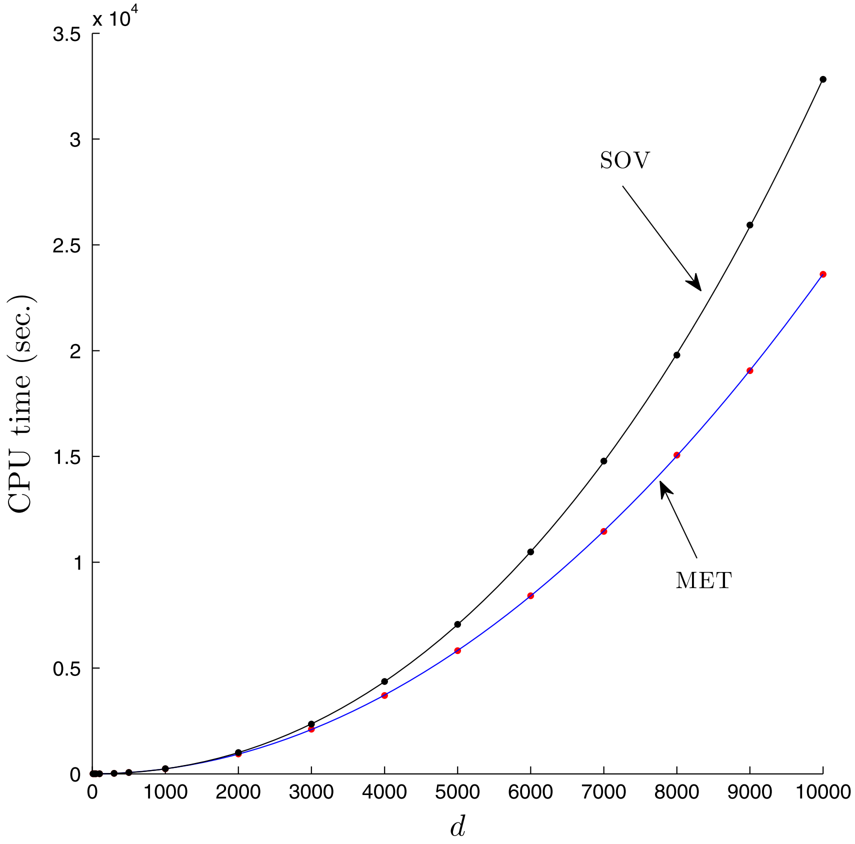

Fig. 3. Graph of computational cost

Table 9. Relative errors

d	*Relative error of MET estimator (%)*	*Relative error of SOV estimator (%)*
10	0.0063	0.0076
30	0.053	0.080
50	0.038	0.090
100	0.15	0.29
300	0.11	1.4
500	0.21	2.0
1000	0.26	3.0
2000	0.18	3.9
3000	0.35	4.8
4000	0.26	8.6
5000	0.33	12
6000	0.28	7.5
7000	0.21	11
8000	0.29	8.3
9000	0.28	15
10000	0.24	12

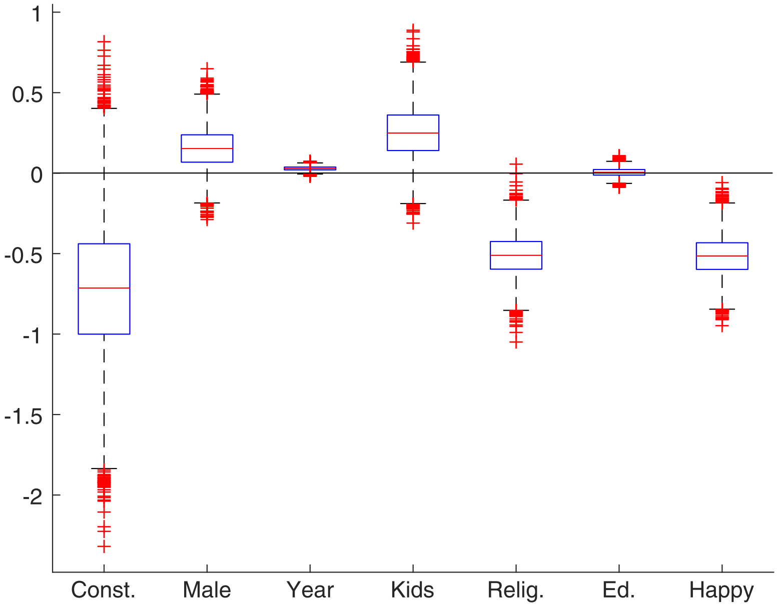

Fig. 4. Marginal distribution of β computed from 8000 exact IID realizations

$$\mathbf{z} = \begin{pmatrix} V^{-1/2}\boldsymbol{\beta} \\ \boldsymbol{\lambda} \end{pmatrix},$$
$$A = (\tilde{X} V^{1/2}, -I),$$
$$\mathbf{l} = \mathbf{0},$$
$$\mathbf{u} = \infty.$$

As a numerical example, we apply the probit model to the widely studied *extramarital affairs* data set from Koop *et al.* (2007). The data set contains $m = 601$ independent observations: the binary response y_i indicates whether the ith respondent has had an extramarital affair; the six explanatory variables ($k = 7$) are a male indicator, number of years married, 'has' or 'has not' children, religious or not, years of formal education and a binary variable denoting whether the marriage is happy or not. Fig. 4 shows the boxplots of the marginal distributions of $\beta_1, \ldots, \beta_7$ based on 8000 IID simulations from the posterior $p(\boldsymbol{\beta}|\mathbf{y})$ with prior covariance $V = 5I$.

The conclusion that only years of marriage, religiosity and conjugal happiness are statistically significant is, of course, well known (Koop *et al.*, 2007) and is used to validate our new simulation scheme. The question is what have we gained in using minimax tilting?

On the one hand, for the first time we have conducted the Bayesian inference by using exact IID samples from the posterior and we did not have to fret about unquantifiable issues such as 'burn-in' and 'mixing speed' as is typical with approximate Markov chain Monte Carlo simulation (Philippe and Robert, 2003).

On the other hand, the acceptance rate in the simulation was $1/217$, i.e. we had to simulate (on average) 217 random vectors to accept one as an exact independent realization from the posterior. Admittedly, this acceptance rate could have been better and as shown in the previous experiments it will deteriorate with increasing dimensionality. However, there are hardly any alternatives for exact sampling—naive acceptance–rejection for the extramarital data would enjoy an acceptance rate of $\mathcal{O}(10^{-146})$ and without minimax tilting (say, with proposal $g(\mathbf{x};\mathbf{0})$) the accept–reject algorithm 2 enjoys an acceptance rate of $\mathcal{O}(10^{-16})$.

Thus, our main point stands: the proposed accept–reject scheme can be used for exact simulation whenever, say, $d \leqslant 100$, and when d is in the thousands it can be used to accelerate Gibbs sampling by grouping or blocking dozens of highly correlated variables together (Chopin, 2011; Philippe and Robert, 2003).

6. Concluding remarks

The minimax tilting method can be effective for exact simulation from the truncated multivariate normal distribution. The method proposed permits us to dispense with Gibbs sampling in dimensions less than 100, and for larger dimensions to accelerate existing Gibbs samplers by sampling jointly hundreds of highly correlated variables.

The minimax approach can also be used to estimate normal probability integrals. Theoretically, the method improves on the already excellent SOV estimator and in a tail asymptotic regime it can achieve the best possible efficiency—VRE. The numerical experiments suggest that the method can be significantly more accurate than the widely used SOV estimator, especially in the tails of the distribution. The experiments also point out its limitations—as the dimensions grow increasingly larger it eventually fails.

The minimax tilting approach in this paper can be extended to other multivariate densities related to the normal distribution. Upcoming work by the author (Botev and L'Ecuyer, 2015) will argue that significant efficiency gains are also possible in the case of the multivariate Student t- and general elliptic distributions for which a strong log-concavity property holds. Just as in the multivariate normal case, the approach permits us to estimate accurately hitherto intractable Student t-probabilities, for which existing estimation schemes exhibit a relative error that is close to 100%.

Acknowledgements

This work was supported by the Australian Research Council under grant DE140100993.

Appendix A

A.1. Proof of lemma 1

First, we show that ψ is a concave function of $\mathbf{x}$ for any $\boldsymbol{\mu}$. To see this, note that, if $Z \sim N(0,1)$ under $\mathbb{P}$, then, by the well-known properties of log-concave measures (Prékopa, 1973) the function $q_1 : \mathbb{R} \to \mathbb{R}$ defined as

$$q_1(w) = \ln\{\mathbb{P}(l \leqslant Z + w \leqslant u)\} = \ln\left[\frac{1}{\sqrt{(2\pi)}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}z^2\right) \mathbb{I}\{(Z+w) \in \mathscr{X}\} \mathrm{d}z\right],$$

where $\mathscr{X} = [l, u]$ is a convex set, is a concave function of $w \in \mathbb{R}$. Hence, for an arbitrary linear map $C \in \mathbb{R}^{d\times 1}$, the function $q_2 : \mathbb{R}^d \to \mathbb{R}$ defined as $q_2(\mathbf{x}) = q_1(C\mathbf{x})$ is concave as well. It follows that each function

$$\ln\{\mathbb{P}(\tilde{l}_k \leqslant Z + \mu_k \leqslant \tilde{u}_k)\} = \ln[\mathbb{P}\{(Z + C_k\mathbf{x}) \in \mathscr{X}_k\}]$$

(using the obvious choices of C_k and $\mathscr{X}_k$) is concave in $\mathbf{x}$. Hence, ψ is concave in $\mathbf{x}$, because it is a non-negative weighted sum of concave functions.

Second, we show that ψ is convex in $\boldsymbol{\mu}$ for each value of $\mathbf{x}$. After some simplification, we can write

$$\psi(\mathbf{x};\boldsymbol{\mu}) = -\mathbf{x}^{\mathrm{T}}\boldsymbol{\mu} + \sum_k \ln(\mathbb{E}[\exp(\mu_k Z)]\, \mathbb{I}(\tilde{l}_k \leqslant Z \leqslant \tilde{u}_k)).$$

Now, each of $\ln(\mathbb{E}[\exp(\mu_k Z)]\mathbb{I}(\tilde{l}_k \leqslant Z \leqslant \tilde{u}_k))$ is convex in μ_k because, up to a normalizing constant, this is the cumulant-generating function of a standard normal random variable Z, truncated to $[\tilde{l}_k, \tilde{u}_k]$. Since a non-negatively weighted sum of convex functions is convex, we conclude that $\psi(\mathbf{x};\boldsymbol{\mu})$ is convex in $\boldsymbol{\mu}$. Finally, since convexity is preserved under pointwise supremum, $\sup_{\mathbf{x}\in\mathscr{C}}\psi(\mathbf{x};\boldsymbol{\mu})$ is still convex in $\boldsymbol{\mu}$. Moreover, here we have the strong min–max property: $\inf_{\mu}\sup_{\mathbf{x}\in\mathscr{C}}\psi(\mathbf{x};\boldsymbol{\mu}) = \sup_{\mathbf{x}\in\mathscr{C}}\inf_{\mu}\psi(\mathbf{x};\boldsymbol{\mu})$, from which lemma 1 follows.

A.2. Proof of theorem 1

Before proceeding with the proof of theorem 1 we note the following results.

First, using the necessary and sufficient condition (13), we can write the solution of problem (12) explicitly as $\mathbf{x}_1 = \gamma L_{11}^{-1}\mathbf{p}_1$ and $\mathbf{x}_2 = \mathbf{0}$ with minimum $(\gamma^2/2)\|L_{11}^{-1}\mathbf{p}_1\|^2$. In addition, from expression (13) we can also deduce that $\boldsymbol{\lambda}_1 = \gamma L_{11}^{-\mathrm{T}} L_{11}^{-1}\mathbf{p}_1 > \mathbf{0}$ and $\mathbf{q} = L_{21}L_{11}^{-1}\mathbf{p}_1 - \mathbf{p}_2 \geqslant \mathbf{0}$.

Second, the asymptotic behaviour of $l(\gamma) = \mathbb{P}(\mathbf{X} \geqslant \gamma\mathbf{l})$ has been established by Hashorva and Hüsler (2003). For convenience, we restate their result by using our simplified notation.

Proposition 1 (Hashorva and Hüsler, 2003). Consider the tail probability $l(\gamma) = \mathbb{P}(\mathbf{X} \geqslant \gamma\mathbf{l})$, where $\mathbf{X} \sim N(\mathbf{0}, \Sigma)$ and $\gamma > 0$ and $\mathbf{l} > \mathbf{0}$. Define the set $\mathscr{J}$ as in expression (14). Then, the tail behaviour of $l(\gamma)$ as $\gamma \uparrow \infty$ is

$$l(\gamma) \simeq c\exp\left[-\frac{\gamma^2}{2}\|L_{11}^{-1}\mathbf{p}_1\|^2 - \sum_{k=1}^{d_1}\ln(\gamma\{L_{11}^{-\mathrm{T}}L_{11}^{-1}\mathbf{p}_1\}_k)\right],$$

where the constant c is given by

$$c = \frac{\mathbb{P}(Y_j > 0, \forall j \in \mathscr{J})}{(2\pi)^{d_1/2}|L_{11}|}, \qquad (Y_1, \ldots, Y_{d_2})^{\mathrm{T}} \sim N(\mathbf{0}, L_{22}L_{22}^{\mathrm{T}})$$

if $\mathscr{J} \neq \emptyset$, and $c = (2\pi)^{-d_1/2}|L_{11}|^{-1}$ if $\mathscr{J} = \emptyset$.

The last two observations pave the way to proving that, depending on the set $\mathscr{J}$, either $\exp\{\psi(\mathbf{x}^*, \boldsymbol{\mu}^*)\} = \mathcal{O}\{l(\gamma)\}$, or $\exp\{\psi(\mathbf{x}^*, \boldsymbol{\mu}^*)\} \simeq l(\gamma)$. The details of the argument are as follows.

In the setting of theorem 1, the Karush–Kuhn–Tucker conditions (10) simplify to

$$\left.\begin{array}{c} \boldsymbol{\mu} - \mathbf{x} + \boldsymbol{\Psi} = \mathbf{0}, \\ -\boldsymbol{\mu} + (\check{L}^{\mathrm{T}} - I)\boldsymbol{\Psi} + \check{L}^{\mathrm{T}}\boldsymbol{\eta} = \mathbf{0}, \\ \boldsymbol{\eta} \geqslant \mathbf{0}, \qquad \gamma\mathbf{p} - L\mathbf{x} \leqslant \mathbf{0}, \\ \boldsymbol{\eta}^{\mathrm{T}}(\gamma\mathbf{p} - L\mathbf{x}) = 0 \end{array}\right\} \tag{17}$$

where $\boldsymbol{\eta}$ is a Lagrange multiplier (corresponding to $\boldsymbol{\eta}_2$ in expression (10)) and we replaced $\mathbf{l}$ with $\gamma\mathbf{p}$.

A.2.1. Case $\mathscr{J} = \emptyset$

We now verify by substitution that, if $\mathscr{J} = \emptyset$, the unique solution of expression (17) is of the asymptotic form

$$\left.\begin{array}{c} \mathbf{x}_1 \simeq \tilde{\mathbf{x}}_1 = \gamma L_{11}^{-1}\mathbf{p}_1, \\ \mathbf{x}_2 \simeq \tilde{\mathbf{x}}_2 = o(\mathbf{1}), \\ \boldsymbol{\mu}_1 \simeq \tilde{\boldsymbol{\mu}}_1 = -\gamma(D_1 L_{11}^{-\mathrm{T}} - I)L_{11}^{-1}\mathbf{p}_1, \\ \boldsymbol{\mu}_2 \simeq \tilde{\boldsymbol{\mu}}_2 = o(\mathbf{1}), \\ \boldsymbol{\eta} \simeq \tilde{\boldsymbol{\eta}} = o(\mathbf{1}). \end{array}\right\} \tag{18}$$

The last equation in expression (17) is obviously satisfied, because $\tilde{\boldsymbol{\eta}} \to 0$ by assumption in expression (18). Next, note that $-\gamma(L_{21}L_{11}^{-1}\mathbf{p}_1 - \mathbf{p}_2) - L_{22}\tilde{\mathbf{x}}_2 = -\gamma\mathbf{q} + o(\mathbf{1}) \downarrow -\infty$, as $\gamma \uparrow \infty$. Hence, the third expression in expression (17) is also satisfied for sufficiently large γ:

$$\gamma\mathbf{p} - L\tilde{\mathbf{x}} = \begin{pmatrix} \gamma\mathbf{p}_1 - L_{11}\tilde{\mathbf{x}}_1 \\ \gamma\mathbf{p}_2 - L_{21}\tilde{\mathbf{x}}_1 - L_{22}\tilde{\mathbf{x}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\gamma(L_{21}L_{11}^{-1}\mathbf{p}_1 - \mathbf{p}_2) - L_{22}\tilde{\mathbf{x}}_2 \end{pmatrix}.$$

Next, note that

$$\begin{array}{c} \tilde{\mathbf{l}}_1 = D_1^{-1}\{\gamma\mathbf{p}_1 - (L_{11} - D_1)\tilde{\mathbf{x}}_1\} = \gamma L_{11}^{-1}\mathbf{p}_1 = \tilde{\mathbf{x}}_1, \\ \tilde{\mathbf{l}}_2 = D_2^{-1}\{\gamma\mathbf{p}_2 - L_{21}\tilde{\mathbf{x}}_1 - (L_{22} - D_2)\tilde{\mathbf{x}}_2\} = -\gamma D_2^{-1}\mathbf{q} + o(\mathbf{1}) \downarrow -\infty. \end{array}$$

Hence from $\tilde{\mathbf{l}}_1-\tilde{\boldsymbol{\mu}}_1=\gamma L_{11}^{-1}\mathbf{p}_1+\gamma(D_1L_{11}^{-\mathrm{T}}-I)L_{11}^{-1}\mathbf{p}_1=\gamma D_1L_{11}^{-\mathrm{T}}L_{11}^{-1}\mathbf{p}_1=D_1\boldsymbol{\lambda}_1>\mathbf{0}$ and $\tilde{\mathbf{l}}_2-\tilde{\boldsymbol{\mu}}_2=-\gamma D_2^{-1}\mathbf{q}+o(\mathbf{1})$, and Mills ratio ($\phi(\gamma;0,1)/\bar{\Phi}(\gamma)\simeq\gamma$ and $\phi(-\gamma;0,1)/\bar{\Phi}(-\gamma)\downarrow 0$) we obtain the asymptotic behaviour of $\boldsymbol{\Psi}$:

$$\begin{aligned}\boldsymbol{\Psi}_1&\simeq\gamma D_1L_{11}^{-\mathrm{T}}L_{11}^{-1}\mathbf{p}_1,\\ \boldsymbol{\Psi}_2&=o(\mathbf{1}),\end{aligned}$$

where we recall that $\boldsymbol{\lambda}_1=\gamma L_{11}^{-\mathrm{T}}L_{11}^{-1}\mathbf{p}_1>\mathbf{0}$. The first equation in expression (17) thus simply verifies that

$$\begin{aligned}\tilde{\mathbf{x}}_1&=\boldsymbol{\Psi}_1+\tilde{\boldsymbol{\mu}}_1\simeq\gamma D_1L_{11}^{-\mathrm{T}}L_{11}^{-1}\mathbf{p}_1-\gamma(D_1L_{11}^{-\mathrm{T}}-I)L_{11}^{-1}\mathbf{p}_1=\gamma L_{11}^{-1}\mathbf{p}_1,\\ \tilde{\mathbf{x}}_2&=\boldsymbol{\Psi}_2+\tilde{\boldsymbol{\mu}}_2=o(\mathbf{1}).\end{aligned}$$

The first and second equations yield $\mathbf{x}=\check{L}^{\mathrm{T}}\boldsymbol{\Psi}=L^{\mathrm{T}}D^{-1}\boldsymbol{\Psi}$, which again is easily verified:

$$\begin{aligned}\mathbf{x}_1&=L_{11}^{\mathrm{T}}D_1^{-1}\boldsymbol{\Psi}_1+L_{21}^{\mathrm{T}}D_2^{-1}\boldsymbol{\Psi}_2\simeq\gamma L_{11}^{-1}\mathbf{p}_1=\tilde{\mathbf{x}}_1,\\ \mathbf{x}_2&=L_{22}^{\mathrm{T}}D_2^{-1}\boldsymbol{\Psi}_2=o(\mathbf{1})=\tilde{\mathbf{x}}_2.\end{aligned}$$

The asymptotic behaviour of $\psi^*=\psi(\tilde{\mathbf{x}};\boldsymbol{\mu}^*)$ is obtained by evaluating ψ at the asymptotic solution (18), i.e.

$$\begin{aligned}\tilde{\psi}\overset{\mathrm{def}}{=}\psi(\tilde{\mathbf{x}};\tilde{\boldsymbol{\mu}})&=\frac{\|\tilde{\boldsymbol{\mu}}\|^2}{2}-\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\boldsymbol{\mu}}+\sum_{k=1}^{d}\ln\{\bar{\Phi}(\tilde{l}_k-\tilde{\mu}_k)\},\qquad\text{where by definition }\bar{\Phi}(x)\overset{\mathrm{def}}{=}\mathbb{P}(Z>x)\\ &=\frac{\|\tilde{\boldsymbol{\mu}}_1\|^2}{2}-\tilde{\mathbf{x}}_1^{\mathrm{T}}\tilde{\boldsymbol{\mu}}_1+\mathcal{O}(\|\tilde{\mathbf{x}}_2\|^2)+\sum_{k=1}^{d_1}\ln\{\bar{\Phi}(\tilde{l}_k-\tilde{\mu}_k)\}+\sum_{k=1}^{d_2}\ln(\bar{\Phi}(-\gamma\{D_2^{-1}\mathbf{q}\}_k+o(1))).\end{aligned}\tag{19}$$

It follows from Mills ratio, $\ln\{\bar{\Phi}(\gamma)\}\simeq-\frac{1}{2}\gamma^2-\ln(\gamma)-\frac{1}{2}\ln(2\pi)$, and $\ln\{\bar{\Phi}(-\gamma)\}\uparrow 0$, that

$$\begin{aligned}\tilde{\psi}&=\frac{\|\tilde{\boldsymbol{\mu}}_1\|^2}{2}-\tilde{\mathbf{x}}_1^{\mathrm{T}}\tilde{\boldsymbol{\mu}}_1-\frac{\gamma^2}{2}\|D_1L_{11}^{-1}L_{11}^{-1}\mathbf{p}_1\|^2-\frac{d_1}{2}\ln(2\pi)-\sum_{k=1}^{d_1}\ln(\gamma\{D_1L_{11}^{-1}L_{11}^{-1}\mathbf{p}_1\}_k)+o(1)\\ &=-\frac{\gamma^2}{2}\|L_{11}^{-1}\mathbf{p}_1\|^2-\frac{d_1}{2}\ln(2\pi)-\ln|L_{11}|-\sum_{k=1}^{d_1}\ln(\gamma\{L_{11}^{-1}L_{11}^{-1}\mathbf{p}_1\}_k)+o(1).\end{aligned}$$

In other words, from proposition 1 we have that $\exp(\tilde{\psi})\simeq l(\gamma)$ as $\gamma\uparrow\infty$. Therefore,

$$\begin{aligned}\frac{\mathrm{var}_{\mu^*}(\hat{l})}{l^2}=\frac{\mathbb{E}_{\mu^*}[\exp\{2\psi(\mathbf{X};\boldsymbol{\mu}^*)\}]}{l^2}-1&\leqslant\frac{\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}\mathbb{E}_{\mu^*}[\exp\{\psi(\mathbf{X};\boldsymbol{\mu}^*)\}]}{l^2}-1\\ &\leqslant\frac{\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}}{l(\gamma)}-1\simeq\frac{\exp(\tilde{\psi})}{l(\gamma)}-1=o(1).\end{aligned}$$

It follows that for $\mathscr{J}=\emptyset$ the minimax estimator (11) exhibits VRE—the best possible asymptotic tail behaviour.

A.2.2. Case $\mathscr{J}\neq\emptyset$

Recall that $(\check{\mathbf{x}},\check{\boldsymbol{\mu}})$ is the solution of the non-linear system (8), as well as the optimization program (7) without its constraint $\mathbf{x}\in\mathscr{C}$ (note that a reordering of the variables via the permutation matrix P does not change the statement of expressions (7) or (8)). We have $\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\leqslant\psi(\check{\mathbf{x}};\check{\boldsymbol{\mu}})$, because dropping a constraint in the maximization of problem (7) cannot reduce the maximum. As in the case of $\mathscr{J}=\emptyset$, one can then verify via direct substitution that

$$\begin{aligned}\tilde{\mathbf{x}}_1&=\gamma L_{11}^{-1}\mathbf{p}_1,\\ \tilde{\mathbf{x}}_2&=\mathcal{O}(\mathbf{1}),\\ \tilde{\boldsymbol{\mu}}_1&=-\gamma(D_1L_{11}^{-\mathrm{T}}-I)L_{11}^{-1}\mathbf{p}_1,\\ \tilde{\boldsymbol{\mu}}_2&=\mathcal{O}(\mathbf{1})\end{aligned}$$

is the asymptotic form of the solution to expression (8). In other words, $\tilde{\psi}=\psi(\tilde{\mathbf{x}};\tilde{\boldsymbol{\mu}})\simeq\psi(\check{\mathbf{x}};\check{\boldsymbol{\mu}})\geqslant\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)$. Similar manipulations to those in expression (19) lead to $\tilde{\psi}=\mathcal{O}(1)-(\gamma^2/2)\|L_{11}^{-1}\mathbf{p}_1\|^2-d_1\ln(\gamma)$. An examination of proposition 1 when $\mathscr{J} \neq \emptyset$ thus shows that $\exp(\tilde{\psi}) = \mathcal{O}\{l(\gamma)\}$ as $\gamma \uparrow \infty$. In other words, $\hat{l}$ is a bounded relative error estimator for $l(\gamma)$:

$$\frac{\mathrm{var}_{\boldsymbol{\mu}^*}(\hat{l})}{l^2} \leqslant \frac{\exp\{\psi(\mathbf{x}^*; \boldsymbol{\mu}^*)\}}{l(\gamma)} - 1 \leqslant \frac{\exp\{\psi(\breve{\mathbf{x}}; \breve{\boldsymbol{\mu}})\}}{l(\gamma)} - 1 \simeq \frac{\exp\{\psi(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\mu}})\}}{l(\gamma)} - 1 = \mathcal{O}(1).$$

A.3. Proof of theorem 2

In the following proof we use the following multi-dimensional Mills ratio (Savage, 1962):

$$\frac{\mathbb{P}(A\mathbf{Z} > \gamma\Sigma\mathbf{l}^*)}{\phi(\gamma\Sigma\mathbf{l}^*; \mathbf{0}, \Sigma)} \simeq \exp\left\{-\sum_k \ln(\gamma l_k^*)\right\}, \qquad \gamma \uparrow \infty. \tag{20}$$

This is a generalization of the well-known one-dimensional result: $\bar{\Phi}(\gamma)/\phi(\gamma; 0, 1) \simeq 1/\gamma$, $\gamma \uparrow \infty$. As in the proof of theorem 1, we proceed to find the asymptotic solution of the non-linear optimization program (7) by considering the necessary and sufficient Karush–Kuhn–Tucker conditions (10). In the set-up of theorem 2 these conditions simplify to (replacing $\mathbf{l}$ with $\gamma\Sigma\mathbf{l}^*$)

$$\left.\begin{array}{c} \boldsymbol{\mu} - \mathbf{x} + \boldsymbol{\Psi} = \mathbf{0}, \\ -\boldsymbol{\mu} + (\breve{L}^{\mathrm{T}} - I)\boldsymbol{\Psi} + \breve{L}^{\mathrm{T}}\boldsymbol{\eta} = \mathbf{0}, \\ \boldsymbol{\eta} \geqslant \mathbf{0}, \qquad \gamma L L^{\mathrm{T}}\mathbf{l}^* - L\mathbf{x} \leqslant \mathbf{0}, \\ \boldsymbol{\eta}^{\mathrm{T}}(\gamma L L^{\mathrm{T}}\mathbf{l}^* - L\mathbf{x}) = 0. \end{array}\right\} \tag{21}$$

We can thus verify via direct substitution that

$$\left.\begin{array}{c} \tilde{\mathbf{x}} = \gamma L^{\mathrm{T}}\mathbf{l}^*, \\ \tilde{\boldsymbol{\mu}} = \gamma(L^{\mathrm{T}} - D)\mathbf{l}^*, \\ \tilde{\boldsymbol{\eta}} = o(\mathbf{1}) \end{array}\right\} \tag{22}$$

satisfy equations (21) asymptotically. The third and fourth equations in expression (21) are satisfied, because $\gamma L L^{\mathrm{T}}\mathbf{l}^* - L\tilde{\mathbf{x}} = \gamma L L^{\mathrm{T}}\mathbf{l}^* - L\gamma L^{\mathrm{T}}\mathbf{l}^* = \mathbf{0}$. Let us now examine the first and second equations in expression (21). First, note that from expression (22)

$$\tilde{\mathbf{l}} - \tilde{\boldsymbol{\mu}} = \gamma\breve{L}L^{\mathrm{T}}\mathbf{l}^* - (\breve{L} - I)\tilde{\mathbf{x}} - \tilde{\boldsymbol{\mu}} = \gamma D\mathbf{l}^* > \mathbf{0}$$

and hence from the one-dimensional Mills ratio we have

$$\Psi_k = \frac{\phi(\tilde{l}_k - \tilde{\mu}_k; 0, 1)}{\bar{\Phi}(\tilde{l}_k - \tilde{\mu}_k)} = \frac{\phi(\gamma D_{kk} l_k^*; 0, 1)}{\bar{\Phi}(\gamma D_{kk} l_k^*)} \simeq \gamma D_{kk} l_k^*, \qquad \gamma \uparrow \infty.$$

In other words, $\boldsymbol{\Psi} \simeq \gamma D\mathbf{l}^*$ as $\gamma \uparrow \infty$. It follows that for the first equation in expression (21) we obtain

$$\tilde{\boldsymbol{\mu}} - \tilde{\mathbf{x}} + \boldsymbol{\Psi} = -\gamma D\mathbf{l}^* + \boldsymbol{\Psi} = o(\mathbf{1})$$

and for the second equation (recall that $\breve{L} = D^{-1}L$, so $\breve{L}^{\mathrm{T}} = L^{\mathrm{T}}D^{-1}$)

$$\begin{aligned} -\tilde{\boldsymbol{\mu}} + (\breve{L}^{\mathrm{T}} - I)\boldsymbol{\Psi} + \breve{L}^{\mathrm{T}}\tilde{\boldsymbol{\eta}} &= -\gamma(\breve{L}^{\mathrm{T}} - I)D\mathbf{l}^* + (\breve{L}^{\mathrm{T}} - I)\boldsymbol{\Psi} + \breve{L}^{\mathrm{T}}\tilde{\boldsymbol{\eta}} \\ &= (\breve{L}^{\mathrm{T}} - I)(\boldsymbol{\Psi} - \gamma D\mathbf{l}^*) + o(\mathbf{1}) = o(\mathbf{1}). \end{aligned}$$

Thus, all the equations in expression (21) are satisfied asymptotically and, since expression (21) has a unique solution, we can conclude that $(\mathbf{x}^*, \boldsymbol{\mu}^*) \simeq (\tilde{\mathbf{x}}, \tilde{\boldsymbol{\mu}})$. We now proceed to substitute the pair $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\mu}})$ into $\psi(\mathbf{x}; \boldsymbol{\mu}) = \|\boldsymbol{\mu}\|^2/2 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\mu} + \Sigma_k \ln\{\bar{\Phi}(\tilde{l}_k - \mu_k)\}$. Using the one-dimensional Mills ratio, $\ln\{\bar{\Phi}(\gamma)\} \simeq -\frac{1}{2}\gamma^2 - \ln(\gamma) - \frac{1}{2}\ln(2\pi)$, we obtain

$$\sum_k \ln\{\bar{\Phi}(\gamma D_{kk} l_k^*)\} \simeq -\frac{\gamma^2}{2}\|D\mathbf{l}^*\|^2 - \sum_k \ln(\gamma D_{kk} l_k^*) - \frac{d}{2}\ln(2\pi), \qquad \gamma \uparrow \infty.$$

As a consequence, using the fact that $\ln|\det(L)|=\Sigma_k\ln(D_{kk})$ (recall that L is triangular with positive diagonal elements), we have

$$\begin{aligned}\tilde{\psi}=\psi(\tilde{\mathbf{x}};\tilde{\boldsymbol{\mu}})&=\psi\{\gamma L^{\mathrm{T}}\mathbf{l}^*;\gamma(L^{\mathrm{T}}-D)\mathbf{l}^*\}\\&=-\frac{1}{2}\|\tilde{\mathbf{x}}\|^2+\frac{\gamma^2}{2}\|D\mathbf{l}^*\|^2+\sum_k\ln\{\bar{\Phi}(\gamma D_{kk}l_k^*)\}\\&\simeq-\frac{\gamma^2}{2}(\mathbf{l}^*)^{\mathrm{T}}LL^{\mathrm{T}}\mathbf{l}^*-\frac{d}{2}\ln(2\pi)-\ln|\det(L)|-\sum_k\ln(\gamma l_k^*).\end{aligned}$$

In other words,

$$\exp\{\psi(\tilde{\mathbf{x}};\tilde{\boldsymbol{\mu}})\}\simeq\phi(\gamma\Sigma\mathbf{l}^*;\mathbf{0},\Sigma)\exp\{-\sum_k\ln(\gamma l_k^*)\},\qquad\gamma\uparrow\infty.$$

However, by the Mills ratio (20), we also have

$$\mathbb{P}(A\mathbf{Z}\geqslant\gamma\Sigma\mathbf{l}^*)\simeq\phi(\gamma\Sigma\mathbf{l}^*;\mathbf{0},\Sigma)\exp\{-\sum_k\ln(\gamma l_k^*)\},\qquad\gamma\uparrow\infty.$$

It follows that $\exp\{\psi(\tilde{\mathbf{x}};\tilde{\boldsymbol{\mu}})\}\simeq l(\gamma)$ and the minimax estimator (11) exhibits VRE:

$$\begin{aligned}\frac{\mathrm{var}_{\mu^*}(\hat{l})}{l^2}=\frac{\mathbb{E}_{\mu^*}[\exp\{2\,\psi(\mathbf{X};\boldsymbol{\mu}^*)\}]}{l^2}-1&\leqslant\frac{\exp\{\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}}{l}-1\\&\simeq\frac{\exp\{\psi(\tilde{\mathbf{x}};\tilde{\boldsymbol{\mu}})\}}{l(\gamma)}-1=o(1),\qquad\gamma\uparrow\infty.\end{aligned}$$

In contrast, for the SOV estimator $\mathring{l}$ we have at most bounded relative error under quite stringent conditions. First, the second moment on the SOV estimator satisfies

$$\liminf_{\gamma\uparrow\infty}\mathbb{E}_{\mathbf{0}}[\exp\{2\,\psi(\mathbf{X};\mathbf{0})\}\geqslant\mathbb{E}_{\mathbf{0}}[\liminf_{\gamma\uparrow\infty}\exp\{2\psi(\mathbf{X};\mathbf{0})\}]$$

and in considering the asymptotics of $\psi(\mathbf{x};\mathbf{0})$ we are free to select $\mathbf{x}$ to obtain the best error behaviour subject to the constraint $\breve{L}\mathbf{x}\geqslant\gamma\breve{L}L^{\mathrm{T}}\mathbf{l}^*$. This gives

$$\exp\{2\,\psi(\mathbf{x};\mathbf{0})\}\simeq\exp\{2\,\psi(\gamma L^{\mathrm{T}}\mathbf{l}^*;\mathbf{0})\}\simeq\frac{1}{\gamma^{2\mathrm{tr}(\Lambda)}}\exp\{-\gamma^2(\mathbf{l}^*)^{\mathrm{T}}L\Lambda L^{\mathrm{T}}\mathbf{l}^*-2c_1\},$$

where $\Lambda=\mathrm{diag}\{(e_1,\ldots,e_d)\}$ is a diagonal matrix such that $e_i=\mathbb{I}(\Sigma_j L_{ji}l_j^*>0)$ and $c_1=\{\mathrm{tr}(\Lambda)/2\}\ln(2\pi)+\Sigma_{k:e_k=1}\ln(\Sigma_j L_{jk}l_j^*)$. It follows that the relative error of the SOV estimator behaves asymptotically as

$$(2\pi)^{d/2}\det(L)\gamma^{d-\mathrm{tr}(\Lambda)}\exp\Big\{\tfrac{1}{2}\gamma^2(\mathbf{l}^*)^{\mathrm{T}}L(I-\Lambda)L^{\mathrm{T}}\mathbf{l}^*-c_1+\sum_k\ln(l_k^*)\Big\}.$$

A.4. *Proof of corollary 1*

Corollary 1 follows from a Pinsker-type inequality (Devroye and Györfi (1985), page 222, theorem 2) by observing that (the expectation operator $\mathbb{E}$ corresponds to the measure $\mathbb{P}$)

$$\begin{aligned}\sup_{\mathscr{A}}|\mathbb{P}(\mathbf{Z}\in\mathscr{A})-\mathbb{P}_{\mu^*}(\mathbf{Z}\in\mathscr{A})|&=\frac{1}{2}\int|f(\mathbf{z})-g(\mathbf{z};\boldsymbol{\mu}^*)|\,\mathrm{d}\mathbf{z}\\&\leqslant\sqrt{\Big\{1-\exp\Big(-\mathbb{E}\Big[\ln\Big\{\frac{f(\mathbf{Z})}{g(\mathbf{Z};\boldsymbol{\mu}^*)}\Big\}\Big]\Big)\Big\}}\\&\leqslant\surd[1-l(\gamma)\exp\{-\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)\}]\\&\simeq\surd\{1-l(\gamma)\exp(-\tilde{\psi})\}=o(1),\end{aligned}$$

where the last equality follows from $\exp(\tilde{\psi})\simeq l(\gamma)$, which is the case when expression (11) is a VRE estimator.

A.5. Proof of lemma 2

That l_{L} is a variational lower bound follows immediately from Jensen's inequality:

$$\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}\exp\left(-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}\underline{\mathrm{var}}(\mathbf{X})\}-\frac{1}{2}\underline{\mathbb{E}}[\mathbf{X}]^{\mathrm{T}}\Sigma^{-1}\underline{\mathbb{E}}[\mathbf{X}]-\underline{\mathbb{E}}[\ln\{\underline{\phi}(\mathbf{X})\}]\right)=\exp\left(\underline{\mathbb{E}}\left[\ln\left\{\frac{\phi(\mathbf{X};\mathbf{0},\Sigma)}{\underline{\phi}(\mathbf{X})}\right\}\right]\right). \tag{23}$$

If $\alpha_i =^{\mathrm{def}} (l_i-\nu_i)/\sigma_i$, $\beta_i =^{\mathrm{def}} (u_i-\nu_i)/\sigma_i$, $p_i=\bar{\Phi}(\alpha_i)-\bar{\Phi}(\beta_i)$ and $\phi(\cdot)\equiv\phi(\cdot;0,1)$, then all the quantities on the left-hand side are available analytically:

$$\left.\begin{aligned}\underline{\mathbb{E}}[X_i]&=\nu_i+\sigma_i\frac{\phi(\alpha_i)-\phi(\beta_i)}{p_i},\\ \mathrm{tr}\{\Sigma^{-1}\underline{\mathrm{var}}(\mathbf{X})\}&=\sum_{i=1}^{d}\{\Sigma^{-1}\}_{i,i}\sigma_i^2\left[1+\frac{\alpha_i\phi(\alpha_i)-\beta_i\phi(\beta_i)}{p_i}-\left\{\frac{\phi(\alpha_i)-\phi(\beta_i)}{p_i}\right\}^2\right],\\ -\underline{\mathbb{E}}[\ln\{\underline{\phi}(\mathbf{X})\}]&=\sum_{i=1}^{d}\frac{\alpha_i\,\phi(\alpha_i)-\beta_i\,\phi(\beta_i)}{2p_i}+\ln[\surd\{2\pi\exp(1)\}\,\sigma_i\,p_i].\end{aligned}\right\} \tag{24}$$

Next, we establish the asymptotic behaviour of $l_{\mathrm{L}}(\gamma)$ under the conditions of theorem 2. Suppose that the pair $(\tilde{\boldsymbol{\nu}},\tilde{\boldsymbol{\sigma}})$ satisfies $\mathrm{diag}^2(\tilde{\boldsymbol{\sigma}})\simeq\Sigma$ and $\tilde{\boldsymbol{\nu}}\simeq\mathbf{l}-\gamma\,\mathrm{diag}^2(\tilde{\boldsymbol{\sigma}})\mathbf{l}^*=\gamma\{\Sigma-\mathrm{diag}^2(\tilde{\boldsymbol{\sigma}})\}\mathbf{l}^*$ as $\gamma\uparrow\infty$. Then, $\boldsymbol{\alpha}\simeq\gamma\,\mathrm{diag}(\tilde{\boldsymbol{\sigma}})\mathbf{l}^*$, which, in combination with $\ln\{\bar{\Phi}(\gamma)\}\simeq-\frac{1}{2}\gamma^2-\ln(\gamma)-\frac{1}{2}\ln(2\pi)$, implies that $\underline{\mathbb{E}}[\mathbf{X}]\simeq\gamma\Sigma\mathbf{l}^*$. Hence, substituting $(\tilde{\boldsymbol{\nu}},\tilde{\boldsymbol{\sigma}})$ into equations (24) and then into the left-hand side of equation (23), and simplifying, we obtain

$$\begin{aligned}l(\gamma)\geqslant l_{\mathrm{L}}\geqslant&\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}\exp\left(-\frac{1}{2}\underline{\mathbb{E}}[\mathbf{X}]^{\mathrm{T}}\Sigma^{-1}\underline{\mathbb{E}}[\mathbf{X}]+\frac{1}{2}\sum_i\left\{\frac{\phi(\alpha_i)}{\bar{\Phi}(\alpha_i)}\right\}^2+\sum_i\ln\{\surd(2\pi)\,\tilde{\sigma}_i\bar{\Phi}(\alpha_i)\}\right)\\ \simeq&\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}\exp\left\{-\frac{1}{2}(\gamma\Sigma\mathbf{l}^*)^{\mathrm{T}}\Sigma^{-1}(\gamma\Sigma\mathbf{l}^*)-\sum_i\ln\left(\frac{\alpha_i}{\tilde{\sigma}_i}\right)\right\}\\ \simeq&\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}\exp\left\{-\frac{\gamma^2}{2}(\mathbf{l}^*)^{\mathrm{T}}\Sigma\mathbf{l}^*-\sum_i\ln(\gamma l_i^*)\right\}\simeq l,\qquad\qquad\gamma\uparrow\infty,\end{aligned}$$

where the last asymptotic equivalence follows from expression (20). Finally, the convergence of expression (16) follows by applying the Pinsker-type inequality (Devroye and Györfi, 1985) in conjunction with

$$\sqrt{\left\{1-\exp\left(-\underline{\mathbb{E}}\left[\ln\left\{\frac{\underline{\phi}(\mathbf{X})}{f(\mathbf{X})}\right\}\right]\right)\right\}}=\sqrt{\left\{1-\frac{1}{l}\exp\left(\underline{\mathbb{E}}\left[\ln\left\{\frac{\phi(\mathbf{X};\mathbf{0},\Sigma)}{\underline{\phi}(\mathbf{X})}\right\}\right]\right)\right\}}\leqslant\sqrt{\left(1-\frac{l_{\mathrm{L}}}{l}\right)}=o(1).$$

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Azaïs, J.-M., Bercu, S., Fort, J. C., Lagnoux, A. and Lé, P. (2010) Simultaneous confidence bands in curve prediction applied to load curves. *Appl. Statist.*, **59**, 889–904.

Bolin, D. and Lindgren, F. (2015) Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Statist. Soc.* B, **77**, 85–106.

Botev, Z. I. and L'Ecuyer, P. (2015) Efficient estimation and simulation of the truncated multivariate Student-t distribution. In *Proc. Winter Simulation Conf., Huntington Beach, Dec. 6th–8th.* To be published.

Botev, Z. I., L'Ecuyer, P. and Tuffin, B. (2013) Markov chain importance sampling with applications to rare event probability estimation. *Statist. Comput.*, **23**, 271–285.

Botts, C. (2013) An accept-reject algorithm for the positive multivariate normal distribution. *Computnl Statist.*, **28**, 1749–1773.

Chopin, N. (2011) Fast simulation of truncated Gaussian distributions. *Statist. Comput.*, **21**, 275–288.

Craig, P. (2008) A new reconstruction of multivariate normal orthant probabilities. *J. R. Statist. Soc.* B, **70**, 227–243.

Davies, P. I. and Higham, N. J. (2000) Numerically stable generation of correlation matrices and their factors. *BIT Numer. Math.*, **40**, 640–651.

Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: the L1 View*. New York: Wiley.

Fernández, P. J., Ferrari, P. A. and Grynberg, S. P. (2007) Perfectly random sampling of truncated multinormal distributions. *Adv. Appl. Probab.*, **39**, 973–990.

Galton, F. (1889) *Natural Inheritance*. London: Macmillan.

Gassmann, H. I. (2003) Multivariate normal probabilities: implementing an old idea of Plackett's. *J. Computnl Graph. Statist.*, **12**, 731–752.

Gassmann, H., Deák, I. and Szántai, T. (2002) Computing multivariate normal probabilities: a new look. *J. Computnl Graph. Statist.*, **11**, 920–949.

Genton, M. G., Ma, Y. and Sang, H. (2011) On the likelihood function of Gaussian max-stable processes. *Biometrika*, **98**, 481–488.

Genz, A. (1992) Numerical computation of multivariate normal probabilities. *J. Computnl Graph. Statist.*, **1**, 141–149.

Genz, A. (2004) Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statist. Comput.*, **14**, 251–260.

Genz, A. and Bretz, F. (2002) Comparison of methods for the computation of multivariate t probabilities. *J. Computnl Graph. Statist.*, **11**, 950–971.

Genz, A. and Bretz, F. (2009) *Computation of Multivariate Normal and t Probabilities*. New York: Springer.

Gerber, M. and Chopin, N. (2015) Sequential quasi Monte Carlo (with discussion). *J. R. Statist. Soc.* B, **77**, 509–579.

Geweke, J. (1991) Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing Science and Statistics: Proc. 23rd Symp. Interface*, pp. 571–578. Fairfax Station: Interface Foundation of North America.

Grün, B. and Hornik, K. (2012) Modelling human immunodeficiency virus ribonucleic acid levels with finite mixtures for censored longitudinal data. *Appl. Statist.*, **61**, 201–218.

Hajivassiliou, V. A. and McFadden, D. L. (1998) The method of simulated scores for the estimation of LDV models. *Econometrica*, **66**, 863–896.

Hashorva, E. and Hüsler, J. (2003) On multivariate Gaussian tails. *Ann. Inst. Statist. Math.*, **55**, 507–522.

Hayter, A. J. and Lin, Y. (2012) The evaluation of two-sided orthant probabilities for a quadrivariate normal distribution. *Computnl Statist.*, **27**, 459–471.

Hayter, A. J. and Lin, Y. (2013) The evaluation of trivariate normal probabilities defined by linear inequalities. *J. Statist. Computn Simuln*, **83**, 668–676.

Huser, R. and Davison, A. C. (2013) Composite likelihood estimation for the Brown–Resnick process. *Biometrika*, **100**, 511–518.

Joe, H. (1995) Approximations to multivariate normal rectangle probabilities based on conditional expectations. *J. Am. Statist. Ass.*, **90**, 957–964.

Koop, G., Poirier, D. J. and Tobias, J. L. (2007) *Bayesian Econometric Methods*. Cambridge: Cambridge University Press.

Kroese, D. P., Taimre, T. and Botev, Z. I. (2011) *Handbook of Monte Carlo Methods*. Hoboken: Wiley.

L'Ecuyer, P., Blanchet, J. H., Tuffin, B. and Glynn, P. W. (2010) Asymptotic robustness of estimators in rare-event simulation. *ACM Trans. Modlng Comput. Simuln*, **20**, article 6.

Miwa, T., Hayter, A. J. and Kuriki, S. (2003) The evaluation of general non-centred orthant probabilities. *J. R. Statist. Soc.* B, **65**, 223–234.

Nomura, N. (2014) Computation of multivariate normal probabilities with polar coordinate systems. *J. Statist. Computn Simuln*, **84**, 491–512.

Nomura, N. (2016) Evaluation of Gaussian orthant probabilities based on orthogonal projections to subspaces. *Statist. Comput.*, **26**, 187–197.

Philippe, A. and Robert, C. P. (2003) Perfect simulation of positive Gaussian distributions. *Statist. Comput.*, **13**, 179–186.

Powell, M. J. D. (1970) A hybrid method for nonlinear equations. *Numer. Meth. Nonlin. Alg. Equns*, **7**, 87–114.

Prékopa, A. (1973) On logarithmic concave measures and functions. *Acta Scient. Math.*, **34**, 335–343.

Sándor, Z. and András, P. (2004) Alternative sampling methods for estimating multivariate normal probabilities. *J. Econmetr.*, **120**, 207–234.

Savage, I. R. (1962) Mills' ratio for multivariate normal distributions. *J. Res. Natn. Bur. Stand.* B, **66**, 93–96.

Tuffin, B. (1999) Bounded normal approximation in simulations of highly reliable markovian systems. *J. Appl. Probab.*, **36**, 974–986.

Vijverberg, W. (1997) Monte Carlo evaluation of multivariate normal probabilities. *J. Econmetr.*, **76**, 281–307.

Wadsworth, J. L. and Tawn, J. A. (2014) Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika*, **101**, 1–15.