The Normal Law Under Linear Restrictions: Simulation and Estimation via Minimax Tilting

Z. I. Botev
The University of New South Wales, botev@unsw.edu.au

Abstract

Simulation from the truncated multivariate normal distribution in high dimensions is a recurrent problem in statistical computing, and is typically only feasible using approximate MCMC sampling. In this article we propose a minimax tilting method for exact iid 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 possesses a rare vanishing relative error asymptotic property. Numerical experiments suggest that the proposed scheme is accurate in a wide range of setups for which competing estimation schemes fail. We give an application to exact iid simulation from the Bayesian posterior of the probit regression model.

1 Introduction

More than a century ago Francis Galton (1889) observed that he scarcely knows "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 article we address some hitherto intractable computational problems related to the d-dimensional multivariate normal law under linear restrictions:

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

where $\mathbb{I}\{\cdot\}$ is the indicator function, $\text{rank}(A) = m \leqslant d$, and $\ell = \mathbb{P}(\mathbf{l} \leqslant A\mathbf{Z} \leqslant \mathbf{u})$ is the probability that a random vector $\mathbf{Z}$ with standard normal distribution in d-dimensions (that is, $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, I_d)$) 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 spacial 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 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 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 (2014b) consider approximation methods for special cases (orthant, bivariate, or trivariate probabilities) and Geweke (1991); Genz (1992); Joe (1995); Vijverberg (1997); Sándor and András (2004); Nomura (2014a) consider estimation schemes applicable for general ℓ. Extensive comparisons amongst the numerous proposals in the literature (Genz and Bretz, 2009; Gassmann et al., 2002; Genz and Bretz, 2002) indicate 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 favorably with highly specialized routines for orthant probabilities (Miwa et al., 2003; Craig, 2008). For this reason, Genz' method is the default choice across different software platforms like Fortran, MATLAB® and **R**.

One of the goals of this article is to propose a new methodology, which not only yields an unbiased estimator 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 error close to 100%) [1].

The obverse to the problem of estimating ℓ is simulation from the truncated multivariate normal $f(\mathbf{z})$. Despite the close relation 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 article is to provide an exact accept-reject sampling scheme for simulation from $f(\mathbf{z})$ in high dimensions, which traditionally calls for approximate MCMC 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 proposed in this article enjoys high acceptance rates in over one hundred 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 the minimax tilting ensures that the new estimator enjoys theoretically better variance properties than the Genz estimator. In an appropriate asymptotic tail regime, the minimax tilting yields an estimator with vanishing relative error (VRE) property (Kroese et al., 2011). Within the light-tailed exponential family, Monte Carlo estimators rarely possess the valuable VRE property (L'Ecuyer et al., 2010) and as yet 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 article we focus on the multivariate normal law due to its central position in statistics, but the proposed methodology can be easily generalized to other multivariate elliptic distributions.

[1] MATLAB® and **R** implementations are available from MATLAB® CENTRAL, `http://www.mathworks.com/matlabcentral/fileexchange/53796`, and the CRAN repository (under the name `TruncatedNormal`), as well as from the author's website: `http://web.maths.unsw.edu.au/~zdravkobotev/`

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 $A = LQ^\top$ be the LQ decomposition of the matrix A, where L is $m \times d$ lower triangular with nonnegative entries down the main diagonal and $Q^\top = Q^{-1}$ is $d \times d$ orthonormal. A simple change of variable $\mathbf{x} \leftarrow Q^\top \mathbf{z}$ then yields:

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

where $\phi(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ denotes the pdf of the $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$ 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 the experimental section. Genz (1992) decomposes 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 separation of variables estimator of ℓ

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

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

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

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

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

Denoting by $\Phi(\cdot)$ the cdf of the standard normal distribution, this gives the following.

Algorithm 2.1 (SOV estimator)

Require: *The lower triangular L such that $A = LQ^\top$, bounds $\mathbf{l}$, $\mathbf{u}$, and uniform sequence $U_1, \ldots, U_{d-1} \overset{iid}{\sim} \mathsf{U}(0, 1)$.*

for $k = 1, 2, \ldots, d-1$ ***do***

Simulate $X_k \sim \mathsf{N}(0, 1)$ conditional 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. That is, set

$$X_k = \Phi^{-1}\left(\Phi(\tilde{l}_k) + U_k\left(\Phi(\tilde{u}_k) - \Phi(\tilde{l}_k)\right)\right).$$

return $\mathring{\ell} = \prod_{k=1}^{d} \left[\Phi(\tilde{u}_k(X_1, \ldots, X_{k-1})) - \Phi(\tilde{l}_k(X_1, \ldots, X_{k-1}))\right]$.

The algorithm can be repeated n times to obtain the iid sample $\mathring{\ell}_1, \ldots, \mathring{\ell}_n$ used for the construction of the unbiased point estimator $\bar{\ell} = (\mathring{\ell}_1 + \cdots + \mathring{\ell}_n)/n$ and its approximate 95% confidence interval $(\bar{\ell} \pm 1.96 \times S/\sqrt{n})$, where S is the sample standard deviation of $\mathring{\ell}_1, \ldots, \mathring{\ell}_n$.

2.1 Variance Reduction via Variable Reordering

Genz and Bretz (2009) suggest 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 $P(1, \ldots, d)^\top = \boldsymbol{\pi}$. It is clear that for any $\boldsymbol{\pi}$ we have $\ell = \mathbb{P}(P\mathbf{l} \leqslant PA\mathbf{Z} \leqslant P\mathbf{u})$. Hence, to estimate ℓ, one can input in the SOV Algorithm 2.1 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 $\mathring{\ell}(\boldsymbol{\pi})$ 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 $\boldsymbol{\pi}^*$ amongst all possible permutations so that

$$\boldsymbol{\pi}^* = \operatorname*{argmin}_{\boldsymbol{\pi}} \operatorname{Var}(\mathring{\ell}(\boldsymbol{\pi}))$$

This is an intractable combinatorial optimization problem whose objective function is not even available. Nevertheless, Genz and Bretz (2009) propose 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 2.1 so that the matrix A and the bounds $\mathbf{l}$ and $\mathbf{u}$ are already in permuted form. We will revisit variable reordering in the numerical experiments in Section 5.

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

2.2 Accept-Reject Simulation

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

Algorithm 2.2 (Accept-Reject Simulation from f)

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

Of course, the accept-reject scheme will only be usable if the probability of acceptance $\mathbb{P}_{\mathbf{0}}(cU \leqslant \phi(\mathbf{X}; \mathbf{0}, I)/g(\mathbf{X})) = \ell/c$ is high and simulation from g is fast. Thus,

this scheme presents two significant challenges which need resolution. The first one 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), while excellent in one and two dimensions, ultimately have acceptance rates of the order $\mathcal{O}(2^{1-d})$ 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 $\exp(\boldsymbol{\mu}^\top \mathbf{x} - K(\boldsymbol{\mu}))\, \phi(\mathbf{x}; \mathbf{0}, I) = \phi(\mathbf{x}; \boldsymbol{\mu}, I)$. Similarly, the tilted version of (4) yields

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

To simplify the notation in the subsequent analysis, let

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

Then, the tilted version of estimator (2) can be written as $\widehat{\ell} = \exp(\psi(\mathbf{X}; \boldsymbol{\mu}))$ with $\mathbf{X} \sim \mathbb{P}_{\boldsymbol{\mu}}$, where $\mathbb{P}_{\boldsymbol{\mu}}$ is the measure with pdf $g(\mathbf{x}; \boldsymbol{\mu}) \stackrel{\text{def}}{=} \prod_{k=1}^d g_k(x_k; \mu_k \,|\, x_1, \ldots, x_{k-1})$. It is now clear that the statistical properties of $\widehat{\ell}$ 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 $\widehat{\ell}$ in a tail asymptotic regime where $\ell \downarrow 0$ 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 $\widehat{\ell}$ is said to be *logarithmically or weakly efficient* if

$$\liminf_{\ell \downarrow 0} \frac{\ln \mathrm{Var}(\widehat{\ell})}{\ln \ell^2} \geqslant 1$$

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

$$\limsup_{\ell \downarrow 0} \frac{\mathrm{Var}(\widehat{\ell})}{\widehat{\ell}^2} \leqslant \text{const.} < \infty.$$

Finally, the best one can hope for in an asymptotic regime is the highly desirable *vanishing relative error* (VRE) property:

$$\limsup_{\ell \downarrow 0} \frac{\mathrm{Var}(\widehat{\ell})}{\widehat{\ell}^2} = 0 \ .$$

An estimator is *strongly efficient* if it exhibits either bounded relative error or VRE. In order 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 $\mathrm{Var}(\widehat{\ell})/\ell^2$, whose behavior is then controlled using the tilting parameter. The strongest type of efficiency VRE is uncommon for light-tailed probabilities, and is typically only achieved within a state-dependent importance sampling framework (L'Ecuyer et al., 2010).

Here we take a different tack, one that exploits features unique to the problem at hand and that will yield efficiency gains in both an asymptotic and non-asymptotic regime. A key result in this direction is the following Lemma 3.1, whose proof is given in the appendix.

Lemma 3.1 (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:

$$\begin{aligned} (\mathbf{x}^*, \boldsymbol{\mu}^*) &= \operatorname*{argmax}_{\mathbf{x}, \boldsymbol{\mu}} \psi(\mathbf{x}; \boldsymbol{\mu}) \\ \textit{subject to: } & \frac{\partial \psi}{\partial \boldsymbol{\mu}} = \mathbf{0}, \qquad \mathbf{x} \in \mathscr{C} \end{aligned} \tag{7}$$

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

Practical Implementation. How do we find the solution of (7) numerically? Without the constraint $\mathbf{x} \in \mathscr{C}$, the solution to (7) would be obtained by solving the nonlinear 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 the case, we introduce the following notation. Let

$D = \text{diag}(L)$, $\check{L} = D^{-1}L$, and

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

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

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

$$\frac{\partial \psi}{\partial \mathbf{x}} = -\boldsymbol{\mu} + (\check{L}^\top - I)\boldsymbol{\Psi} = \mathbf{0}, \qquad \frac{\partial \psi}{\partial \boldsymbol{\mu}} = \boldsymbol{\mu} - \mathbf{x} + \boldsymbol{\Psi} = \mathbf{0}\,, \tag{8}$$

and the Jacobian matrix has elements:

$$\frac{\partial^2 \psi}{\partial \boldsymbol{\mu}^2} = I + \text{diag}\,(\boldsymbol{\Psi}')\,, \quad \frac{\partial^2 \psi}{\partial \boldsymbol{\mu} \partial \mathbf{x}} = (\check{L} - I)\text{diag}(\boldsymbol{\Psi}') - I, \quad \frac{\partial^2 \psi}{\partial \mathbf{x}^2} = (\check{L} - I)^\top \text{diag}\,(\boldsymbol{\Psi}')\,(\check{L} - I)\,. \tag{9}$$

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

$$\begin{aligned} \partial\psi/\partial\boldsymbol{\mu} = \mathbf{0}, \qquad & \partial\psi/\partial\mathbf{x} - \check{L}^\top\boldsymbol{\eta}_1 + \check{L}^\top\boldsymbol{\eta}_2 = \mathbf{0} \\ \boldsymbol{\eta}_1 \geqslant \mathbf{0}, \quad & L\mathbf{x} - \mathbf{u} \leqslant \mathbf{0}, \quad \boldsymbol{\eta}_1^\top(L\mathbf{x} - \mathbf{u}) = \mathbf{0} \\ \boldsymbol{\eta}_2 \geqslant \mathbf{0}, \quad & -L\mathbf{x} + \mathbf{l} \leqslant \mathbf{0}, \quad \boldsymbol{\eta}_2^\top(L\mathbf{x} - \mathbf{l}) = \mathbf{0}, \end{aligned} \tag{10}$$

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

Suppose we find the unique solution of the nonlinear system (8) using, for example, a trust-region Dogleg method (Powell, 1970). If we denote the solution to (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 (8) does not lie in $\mathscr{C}$, then $(\check{\mathbf{x}}; \check{\boldsymbol{\mu}})$ will be suboptimal and, in order to compute $(\mathbf{x}^*; \boldsymbol{\mu}^*)$, one has to use a constrained convex optimization solver. This observation then leads to the following procedure.

Algorithm 3.1 (Computation of optimal pair $(\mathbf{x}^*, \boldsymbol{\mu}^*)$)

Use Powell's (1970) Dogleg method on (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 $(\mathbf{x}^*, \boldsymbol{\mu}^*)$

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

4 Theoretical Properties of Minimax Tilting

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

$$\widehat{\ell} = \exp\left(\psi(\mathbf{X}; \boldsymbol{\mu}^*)\right), \qquad \mathbf{X} \sim \mathbb{P}_{\boldsymbol{\mu}^*}, \tag{11}$$

achieves the best possible efficiency in a tail asymptotic regime.

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

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

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

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:

$$\begin{aligned} \min_{\mathbf{x}} \ & \frac{1}{2}\|\mathbf{x}\|^2 \\ \text{subject to: } & L\mathbf{x} \geqslant \gamma \mathbf{p} \end{aligned} \tag{12}$$

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

$$\begin{aligned} \mathbf{x} - L^\top \boldsymbol{\lambda} &= \mathbf{0} \\ \boldsymbol{\lambda} \geqslant \mathbf{0}, \ \ \gamma \mathbf{p} - L\mathbf{x} &\leqslant \mathbf{0} \\ \boldsymbol{\lambda}^\top(\gamma \mathbf{p} - L\mathbf{x}) &= 0 \,, \end{aligned} \tag{13}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ is a Lagrange multiplier vector. Suppose the number of active constraints in (12) is d_1 and the number of inactive constraints is d_2, where $d_1 + d_2 = d$. Note that 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 $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top$ 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 (13) correspond to $\boldsymbol{\lambda}_1 > \mathbf{0}$ and all the inactive ones to $\boldsymbol{\lambda}_2 = \mathbf{0}$. Similarly, we define a partitioning for $\mathbf{x}, \mathbf{p}$, and the lower triangular

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

Note that the only reason for introducing the above variable reordering via the permutation matrix P and insisting that all active constraints of (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 4.1 in its simplest and neatest form.

Theorem 4.1 (Strong Efficiency of Minimax Estimator) *Consider the estimation of the probability*

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

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

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

and let $\mathscr{J}$ *be the set of indices for which the components of the vector* $\mathbf{q}$ *are zero, that is,*

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

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

$$\limsup_{\gamma \uparrow \infty} \frac{Var_{\mu^*}(\widehat{\ell}(\gamma))}{\ell^2(\gamma)} = 0\ .$$

Alternatively, if $\mathscr{J} \neq \emptyset$*, then* $\widehat{\ell}$ *is a bounded relative error estimator:*

$$\limsup_{\gamma \uparrow \infty} \frac{Var_{\mu^*}(\widehat{\ell}(\gamma))}{\ell^2(\gamma)} < const. < \infty.$$

The theorem 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 $\Sigma = AA^\top$, then we always have VRE.

Theorem 4.2 (Minimax Vanishing Relative Error) *Consider the estimation of the tail probability* $\ell(\gamma) = \mathbb{P}(\gamma \mathbf{l} \leqslant A\mathbf{Z} \leqslant \boldsymbol{\infty})$*, where* $\mathbf{l} = \Sigma \mathbf{l}^*$ *for some positive weight* $\mathbf{l}^* > \mathbf{0}$*. Then, the minimax estimator* (11) *is a vanishing relative error estimator.*

In contrast, under the additional assumption $L^\top \mathbf{l}^* > \mathbf{0}$ *(strong positive covariance), where* L *is the lower triangular factor of* $\Sigma = LL^\top$*, the SOV estimator* $\mathring{\ell}$ *is a bounded relative error estimator; otherwise, it is a divergent one*[2]*:*

$$\frac{Var_{\mathbf{0}}(\exp(\psi(\mathbf{X};\mathbf{0})))}{\ell^2(\gamma)} \simeq \begin{cases} \mathcal{O}(1), & \text{if } L^\top \mathbf{l}^* > \mathbf{0} \\ \exp(\mathcal{O}(\gamma^2) + \mathcal{O}(\ln \gamma) + \mathcal{O}(1)), & \text{otherwise} \end{cases}.$$

Note that the permutation matrix P plays no role in the statement of Theorem 4.2 (we can assume $P = I$), 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 4.1 and 4.2, for the obverse problem of simulation from the truncated multivariate normal, we obtain the following result.

[2]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.

Corollary 4.1 (Asymptotically Efficient Simulation) *Suppose that the instrumental density in the Accept-Reject Algorithm 2.2 for simulation from*

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

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* $\mathbb{P}$*:*

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

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

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 $\ell_{\mathrm{L}} \leqslant \ell$ be a lower bound to ℓ. Then, by Hoeffding's inequality for $\bar{\ell} = (\widehat{\ell_1} + \cdots + \widehat{\ell_n})/n$ with

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

we obtain: $\mathbb{P}_{\boldsymbol{\mu}^*}(\bar{\ell} - \varepsilon \leqslant \ell \leqslant \bar{\ell} + \varepsilon) \geqslant 1 - \alpha$.

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

Lemma 4.1 (Cross Entropy Lower Bound) *Define the product measure* $\underline{\mathbb{P}}$ *with pdf*

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

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

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

where $\Sigma = AA^\top$*. Then,* $\ell_{\mathrm{L}} \leqslant \ell$ *is a variational lower bound to* ℓ*. In addition, under the conditions of Theorem 4.2, namely,* $(\mathbf{l}, \mathbf{u}) = (\gamma \Sigma \mathbf{l}^*, \boldsymbol{\infty})$ *, we have that* $\ell_{\mathrm{L}} \uparrow \ell(\gamma)$ *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 $\mathbb{P}_{\boldsymbol{\mu}^*}$. Unfortunately, despite the similarity of the results in Theorem 4.2 and Lemma 4.1, 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 4.1 is for constructing an exact confidence interval and lower bound to ℓ in the tails of the normal distribution.

Note that under the conditions of Theorem 4.2, the minimax estimator enjoys the bounded normal approximation property (Tuffin, 1999). That is, if $\bar{\ell}$ and S^2 are the mean and sample variance of the iid $\widehat{\ell}_1, \ldots, \widehat{\ell}_n$, and $F_n(x)$ is the empirical cdf of $T_n = \sqrt{n}(\bar{\ell} - \ell)/S$, then we have the Berry–Esséen bound, uniformly in γ:

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

This Berry–Esséen bound implies that the coverage error of the approximate $(1-\alpha)$ level confidence interval $\bar{\ell} \pm z_{1-\alpha/2} \times S/\sqrt{n}$ remains of the order $\mathcal{O}(n^{-1/2})$, even as $\ell \downarrow 0$. Thus, if a lower bound ℓ_{L} is not easily available, one can still rely on the confidence interval 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 a number of test cases 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 $\mathbf{X} = A\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ conditional on $\mathbf{X} \in \mathscr{A}$, and the estimation of ℓ in (1).

In all examples we compare the separation-of-variables (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 (2014a) and the specialized orthant probability algorithm of Miwa et al. (2003); Craig (2008); Nomura (2014b). Unfortunately, unless a special autoregressive 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 $\mathcal{O}(d! \times n)$, 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 MET estimators are smooth, one can seek further gains in efficiency using randomized quasi Monte Carlo. The idea behind quasi Monte Carlo is to reduce the error of the estimator by using *quasirandom* 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 MET estimator we use the n-point Richtmyer quasirandom 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 as follows.

Algorithm 5.1 (Randomized Quasi Monte Carlo (Genz and Bretz, 2009))

Require: *Dimension d and sample size n.*

$d' \leftarrow \lceil 5d \ln(d+1)/4 \rceil, \quad n' \leftarrow \lceil \frac{n}{12} \rceil$

Let $p_1, \ldots, p_{d'}$ be the first d' prime numbers.
$\mathbf{q}_i \leftarrow \sqrt{p_i} \times (1, \ldots, n')^\top$ *for* $i = 1, \ldots, d'$
for $k = 1, \ldots, 12$ ***do***
  for $i = 1, \ldots, d-1$ ***do***
    Let $U \sim \mathsf{U}(0,1)$*, independently.*
    $\mathbf{s}_i \leftarrow |2 \times [(\mathbf{q}_i + U) \mod 1] - 1|$
  $\mathbf{qms} \leftarrow (\mathbf{s}_1, \ldots, \mathbf{s}_{d-1})$
  Use the sequence **qms** *to compute an n'-point sample average estimator $\widehat{\ell}_k$.*
return $\bar{\ell} \leftarrow \frac{1}{12}\sum_k \widehat{\ell}_k$ *with estimated relative error* $\frac{1}{12}\sqrt{\sum_k (\widehat{\ell}_k - \bar{\ell})^2}\Big/\bar{\ell}$.

Note that, since there is no need to integrate the x_d-th component, the loop over i 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 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.

Example I (Fernández et al., 2007). Consider $\mathscr{A} = [1/2, 1]^d$ with a covariance matrix

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

Columns three and four in Table 1 show the estimates of ℓ for various values of d. The brackets give the estimated relative error in percentage.

Figure 1: Estimates of ℓ for various values of d using $n = 10^4$ replications.

d	ℓ_{L}	SOV	MET	$\exp(\psi(\mathbf{x}^*;\boldsymbol{\mu}^*))$	worst err.	accept pr.
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

The second column shows the lower bound discussed in Lemma 4.1 and column five shows the deterministic upper bound. These two bounds can then be used to compute the exact 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 column six under the "worst err." heading)

$$\sqrt{\mathrm{Var}(\bar{\ell})}/\ell \leqslant (\exp(\psi(\mathbf{x}^*;\boldsymbol{\mu}^*)) / \ell_{\mathrm{L}} - 1)/\sqrt{n}.$$

Finally, column seven (accept pr.) gives the acceptance rate of Algorithm 2.2 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 column seven 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 MET estimators gradually deteriorates with increasing d. However, the SOV estimator has larger relative error, does not give meaningful results for $d > 25$, and possesses no theoretical quantification of its performance. In contrast, the MET estimator is guaranteed to have better relative error than the one given in column six (worst err.).

Finally, in further numerical experiments (not displayed here) we observed that the width, ε, of the *exact* confidence interval, $\bar{\ell} \pm \varepsilon$ 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 $\bar{\ell} \pm z_{1-\alpha/2} \times S/\sqrt{n(\varepsilon)}$.

Example II (Fernández et al., 2007). Consider the hypercube $\mathscr{A} = [0,1]^d$ and the isotopic covariance with elements

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

Figure 2: Estimates of ℓ for various values of d using $n = 10^4$ replications.

d	ℓ_{L}	SOV	MET	$\exp(\psi(\mathbf{x}^*;\boldsymbol{\mu}^*))$	worst err.	accept pr.
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

Observe how rapidly the probabilities 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 get larger and larger, making naive accept-reject simulation infeasible. These small probabilities sometimes present not only theoretical challenges (rare-event estimation), but practical ones like representation in finite precision arithmetic and numerical underflow. For instance, in using the SOV estimator Grün and Hornik (2012) note 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 compu-

tation, but consistent with Theorems 4.1 and 4.2, it is typically much more robust than the SOV estimator in estimating small probabilities.

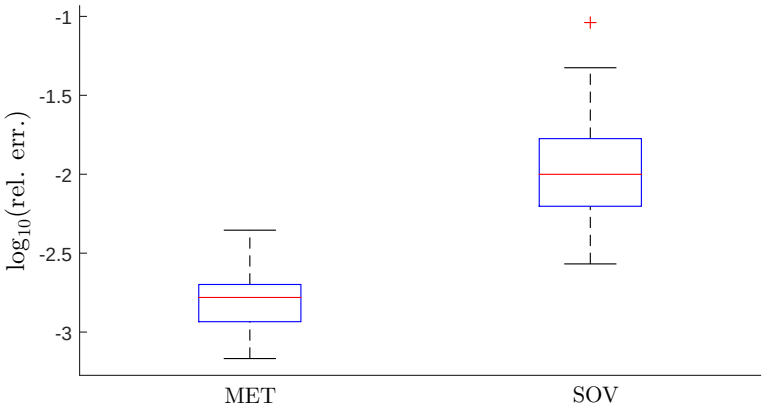
5.2 Random Correlation Matrices

One can argue that the covariance matrices 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) find it desirable to use random correlation matrices. In the subsequent examples we use the method of Davies and Higham (2000) to simulate random test correlation matrices whose eigenvalues are uniformly distributed over the simplex $\{\mathbf{x} : x_1 + \cdots + x_d = d\}$.

Example III. A natural question is whether the MET estimator would still be preferable when integrating over a "non-tail" region such as $\mathscr{A} = [-1/2, \infty]^{100}$. The table below summarizes the output of running the algorithms on 100 independently simulated random correlation matrices. Both the SOV and MET estimators used $n = 10^5$ quasi Monte Carlo points. The 'accept rate' row displays the five number summary of the estimated acceptance probability of Algorithm 2.2.

Figure 3: Table: five number summary for relative error based on 100 independent replications; Graph: boxplots of these 100 outcomes on logarithmic scale.

	min	1-st quartile	median	3-rd quartile	max
MET	0.07%	0.12%	0.17%	0.20%	0.44%
SOV	0.27%	0.63%	1.00%	1.68%	9.14%
accept rate	1.2%	3.9%	5.5%	7.3%	12%



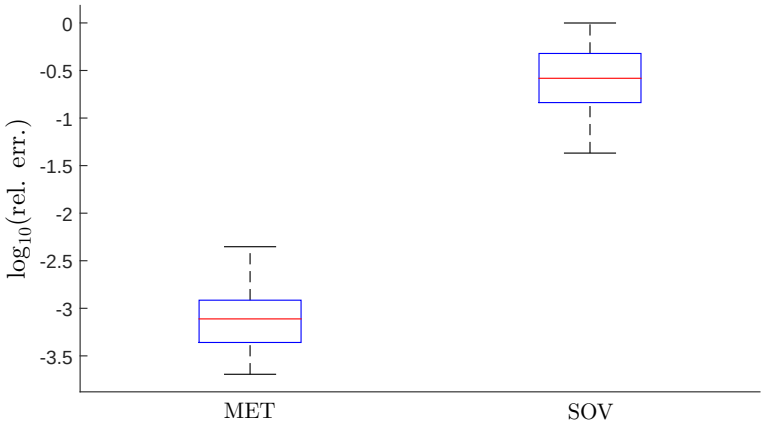
So far we have said little about the cost of computing the optimal pair $(\mathbf{x}^*; \boldsymbol{\mu}^*)$, and the measures of efficiency we have considered do not account for the computational cost of the estimators. The reason for this is that in the examples we investigated, the computing time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

In the current example, the numerical experiments suggest that the MET estimator is roughly 20% more costly than the SOV estimator. If one adjusts the results in Figure 3 in order 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 $1/1.2 \approx 0.83$ percent, which is hardly significant.

Example IV. Finally, we wish to know if the strong efficiency described in Theorem 4.1 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 $n = 10^5$. The following table and graph summarize the results of 100 replications.

Figure 4: Relative errors of SOV and MET estimators over 100 random correlation cases.

	min	1-st quartile	median	3-rd quartile	max
MET	0.020%	0.044%	0.077%	0.12%	0.44%
SOV	4.3%	15%	26%	48%	99%
accept rate	1.5%	10%	18%	26%	43%



As seen from the results, in this particular example the variance of the MET estimator is typically more than 10^5 times smaller than 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 MCMC, it is not a panacea against the curse of dimensionality. The acceptance probability of Algorithm 2.2 ultimately becomes a rare-event probability as the dimensions keep increasing, because the bounded or vanishing relative error properties of $\widehat{\ell}$ 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 ℓ does

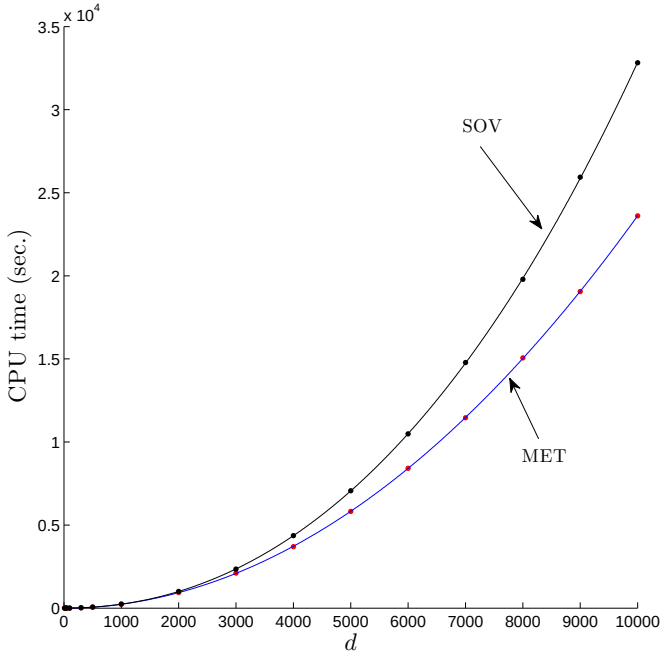
not decay too fast in d. In this regard, Miwa et al. (2003); Craig (2008) study the orthant probability $\ell = \mathbb{P}(\mathbf{X} \in [0,\infty]^d)$ with the positive correlation structure

$$\Sigma = \frac{1}{2} I + \frac{1}{2} \mathbf{1}\mathbf{1}^\top \ .$$

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

Figure 5 shows the output of a numerical experiment with $n = 10^5$ for various values of d. The graph on the left gives the computational cost in seconds. Both the SOV and the MET estimators have cost of $\mathcal{O}(d^3)$ — hence the excellent agreement with the least squares cubic polynomials fitted to the empirical CPU data. The table on the right 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.

Figure 5: Graph: computational cost in seconds; Table: relative error in percentage;



d	MET	SOV
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%

This example confirms the result in Theorem 4.2 that the SOV estimator works better in settings with strongly positive correlation structure (but poorly with negative correlation). Further, the results suggest the MET estimator is also aided by the presence of positive correlation.

5.4 Exact Simulation of Probit Posterior

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

- Prior: $p(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top V^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right)$ with $\boldsymbol{\beta} \in \mathbb{R}^k$ and for simplicity $\boldsymbol{\beta}_0 = \mathbf{0}$;
- Likelihood: $p(\mathbf{y}\,|\,\boldsymbol{\beta}) \propto \exp\left(\sum_{i=1}^m \ln \Phi\big((2y_i-1)\mathbf{x}_i^\top \boldsymbol{\beta}\big)\right)$.

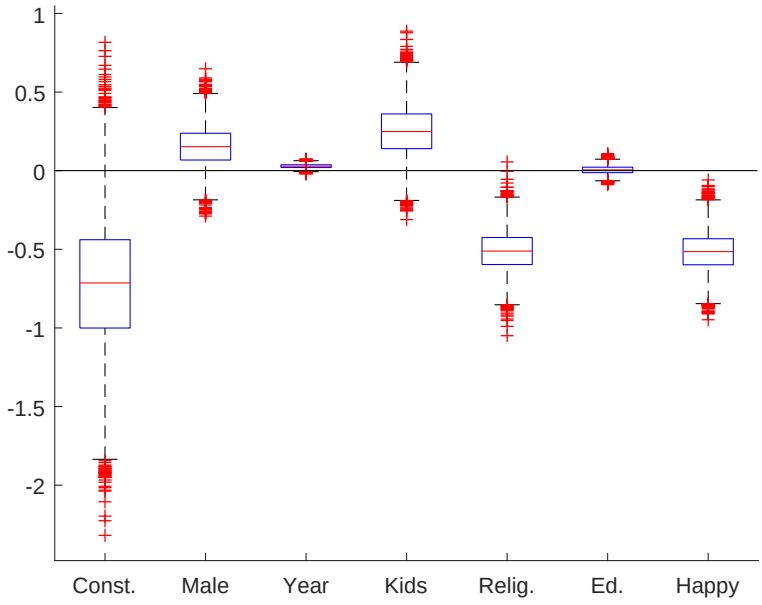
The challenge is to simulate from the posterior $p(\boldsymbol{\beta}\,|\,\mathbf{y})$. One can use latent variables (Albert and Chib, 1993) to represent the posterior as the marginal of a truncated multivariate normal. Let $\boldsymbol{\lambda} \sim \mathsf{N}(0, I_m)$ be latent variables and define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

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

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

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

Figure 6: Marginal distribution of $\boldsymbol{\beta}$ computed from 8000 exact iid realizations.



As an numerical example, we apply the probit model to the widely studied *extramarital affairs* dataset from Koop et al. (2007). The dataset contains $m = 601$ independent observations: the binary response y_i indicates if the i-th respondent has had an extramarital affair; the six explanatory variables ($k = 7$) are male indicator (Male), number of years married (Year), 'has' or 'has not' children (Kids), religious or not (Relig.), years of formal education (Ed.), and a binary variable denoting whether the marriage is happy or not (Happy). Figure 6 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 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 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 MCMC simulation (Philippe and Robert, 2003).

On the other hand, the acceptance rate in the simulation was $1/217$, that is, 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 is going to 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 2.2 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).

Concluding Remarks

The minimax tilting method can be effective for exact simulation from the truncated multivariate normal distribution. The proposed method 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 — vanishing relative error. The numerical experiments suggest that the proposed method can be significantly more accurate than the widely used SOV estimator, especially in the tails of the distribution. The experiments also point out to its limitations — as the dimensions get larger and larger it eventually fails.

The minimax tilting approach in this article can be extended to other multivariate densities related to the normal. Upcoming work by the author 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 relative error close to 100%.

Acknowledgments

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

A Appendix

A.1 Proof of Lemma 3.1

First, we show that ψ is a concave function of $\mathbf{x}$ for any $\boldsymbol{\mu}$. To see this, note that if $Z \sim \mathsf{N}(0,1)$ 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

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

where $\mathscr{Z} = [l, u]$ 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

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

(using the obvious choices of C_k and $\mathscr{Z}_k$) 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

$$\psi(\mathbf{x};\boldsymbol{\mu}) = -\mathbf{x}^\top\boldsymbol{\mu} + \sum_k \ln \mathbb{E}\, \exp\left(\mu_k Z\right) \mathbb{I}_{\{\tilde{l}_k \leqslant Z \leqslant \tilde{u}_k\}} \ .$$

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: $\inf_{\boldsymbol{\mu}} \sup_{\mathbf{x}\in\mathscr{C}} \psi(\mathbf{x};\boldsymbol{\mu}) = \sup_{\mathbf{x}\in\mathscr{C}} \inf_{\boldsymbol{\mu}} \psi(\mathbf{x};\boldsymbol{\mu})$, from which the lemma follows. □

A.2 Proof of Theorem 4.1

Before proceeding with the proof we note the following.

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

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

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

$$\ell(\gamma) \simeq c \times \exp\left(-\frac{\gamma^2}{2}\|L_{11}^{-1}\mathbf{p}_1\|^2 - \sum_{k=1}^{d_1} \ln\left(\gamma\left\{L_{11}^{-\top}L_{11}^{-1}\mathbf{p}_1\right\}_k\right)\right),$$

where the constant c is given by:

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

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 $\exp(\psi(\mathbf{x}^*, \boldsymbol{\mu}^*)) = \mathcal{O}(\ell(\gamma))$, or $\exp(\psi(\mathbf{x}^*, \boldsymbol{\mu}^*)) \simeq \ell(\gamma)$. The details of the argument are as follows.

In the setting of Theorem 4.1, the Karusch-Kuhn-Tucker conditions (10) simplify to:

$$\begin{aligned} \boldsymbol{\mu} - \mathbf{x} + \boldsymbol{\Psi} &= \mathbf{0} \\ -\boldsymbol{\mu} + (\check{L}^\top - I)\boldsymbol{\Psi} + \check{L}^\top\boldsymbol{\eta} &= \mathbf{0} \\ \boldsymbol{\eta} \geqslant \mathbf{0}, \quad \gamma\mathbf{p} - L\mathbf{x} &\leqslant \mathbf{0} \\ \boldsymbol{\eta}^\top(\gamma\mathbf{p} - L\mathbf{x}) &= 0 \end{aligned} \tag{17}$$

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

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

$$\begin{aligned} \mathbf{x}_1 &\simeq \tilde{\mathbf{x}}_1 = \gamma L_{11}^{-1}\mathbf{p}_1 \\ \mathbf{x}_2 &\simeq \tilde{\mathbf{x}}_2 = o(\mathbf{1}) \\ \boldsymbol{\mu}_1 &\simeq \tilde{\boldsymbol{\mu}}_1 = -\gamma(D_1 L_{11}^{-\top} - I)L_{11}^{-1}\mathbf{p}_1 \\ \boldsymbol{\mu}_2 &\simeq \tilde{\boldsymbol{\mu}}_2 = o(\mathbf{1}) \\ \boldsymbol{\eta} &\simeq \tilde{\boldsymbol{\eta}} = o(\mathbf{1}) \end{aligned} \tag{18}$$

Equation four in (17) is obviously satisfied, because $\tilde{\boldsymbol{\eta}}$ tends to zero by assumption in (18). Next, note that $-\gamma\left(L_{21}L_{11}^{-1}\mathbf{p}_1 - \mathbf{p}_2\right) - L_{22}\tilde{\mathbf{x}}_2 = -\gamma\mathbf{q} + o(\mathbf{1}) \downarrow -\infty$, as $\gamma \uparrow \infty$. Hence, line three in (17) is also satisfied for sufficiently large γ:

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

Next, note that

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

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

$$\boldsymbol{\Psi}_1 \simeq \gamma D_1 L_{11}^{-\top}L_{11}^{-1}\mathbf{p}_1, \qquad \boldsymbol{\Psi}_2 = o(\mathbf{1})\ ,$$

where we recall that $\boldsymbol{\lambda}_1 = \gamma L_{11}^{-\top} L_{11}^{-1} \mathbf{p}_1 > \mathbf{0}$. Equation one in (17) thus simply verifies that

$$\begin{aligned}\tilde{\mathbf{x}}_1 &= \boldsymbol{\Psi}_1 + \tilde{\boldsymbol{\mu}}_1 \simeq \gamma D_1 L_{11}^{-\top} L_{11}^{-1} \mathbf{p}_1 - \gamma (D_1 L_{11}^{-\top} - I) L_{11}^{-1} \mathbf{p}_1 = \gamma L_{11}^{-1} \mathbf{p}_1 \\ \tilde{\mathbf{x}}_2 &= \boldsymbol{\Psi}_2 + \tilde{\boldsymbol{\mu}}_2 = o(\mathbf{1})\end{aligned}$$

Equation one and two yield $\mathbf{x} = \check{L}^\top \boldsymbol{\Psi} = L^\top D^{-1} \boldsymbol{\Psi}$, which again is easily verified:

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

The asymptotic behavior of $\psi^* = \psi(\mathbf{x}^*; \boldsymbol{\mu}^*)$ is obtained by evaluating ψ at the asymptotic solution (18), that is, $\tilde{\psi} \stackrel{\text{def}}{=} \psi(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\mu}}) =$

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

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

$$\begin{aligned}\tilde{\psi} &= \frac{\|\tilde{\boldsymbol{\mu}}_1\|^2}{2} - \tilde{\mathbf{x}}_1^\top \tilde{\boldsymbol{\mu}}_1 - \frac{\gamma^2}{2} \|D_1 L_{11}^{-1} L_{11}^{-1} \mathbf{p}_1\|^2 - \frac{d_1}{2} \ln(2\pi) - \sum_{k=1}^{d_1} \ln(\gamma \{D_1 L_{11}^{-1} L_{11}^{-1} \mathbf{p}_1\}_k) + o(1) \\ &= -\frac{\gamma^2}{2} \|L_{11}^{-1} \mathbf{p}_1\|^2 - \frac{d_1}{2} \ln(2\pi) - \ln |L_{11}| - \sum_{k=1}^{d_1} \ln(\gamma \{L_{11}^{-1} L_{11}^{-1} \mathbf{p}_1\}_k) + o(1)\end{aligned}$$

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

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

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

Case $\mathscr{J} \neq \emptyset$. Recall that $(\check{\mathbf{x}}, \check{\boldsymbol{\mu}})$ is the solution of the nonlinear 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 (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 (7) cannot reduce the maximum. As in the case of $\mathscr{J} = \emptyset$, one can then verify via direct substitution that

$$\tilde{\mathbf{x}}_1 = \gamma L_{11}^{-1} \mathbf{p}_1, \quad \tilde{\mathbf{x}}_2 = \mathcal{O}(\mathbf{1}), \quad \tilde{\boldsymbol{\mu}}_1 = -\gamma (D_1 L_{11}^{-\top} - I) L_{11}^{-1} \mathbf{p}_1, \quad \tilde{\boldsymbol{\mu}}_2 = \mathcal{O}(\mathbf{1})$$

is the asymptotic form of the solution to (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 as the ones in (19) lead to $\tilde{\psi} = \mathcal{O}(1) - \frac{\gamma^2}{2}\|L_{11}^{-1}\mathbf{p}_1\|^2 - d_1 \ln \gamma$. An examination of Proposition A.1 when $\mathscr{J} \neq \emptyset$ thus shows that $\exp(\tilde{\psi}) = \mathcal{O}(\ell(\gamma))$ as $\gamma \uparrow \infty$. In other words, $\widehat{\ell}$ is a bounded relative error estimator for $\ell(\gamma)$:

$$\frac{\mathrm{Var}_{\boldsymbol{\mu}^*}(\widehat{\ell})}{\ell^2} \leqslant \frac{\exp(\psi(\mathbf{x}^*;\boldsymbol{\mu}^*))}{\ell(\gamma)} - 1 \leqslant \frac{\exp(\psi(\check{\mathbf{x}};\check{\boldsymbol{\mu}}))}{\ell(\gamma)} - 1 \simeq \frac{\exp(\psi(\tilde{\mathbf{x}};\tilde{\boldsymbol{\mu}}))}{\ell(\gamma)} - 1 = \mathcal{O}(1) \,.$$

A.3 Proof of Theorem 4.2

In the following proof we use the following multidimensional Mill's ratio (Savage, 1962):

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

This is a generalization of the well-known one-dimensional result: $\frac{\overline{\Phi}(\gamma)}{\phi(\gamma;0,1)} \simeq \frac{1}{\gamma}$, $\gamma \uparrow \infty$. As in the proof of Theorem 4.1, we proceed to find the asymptotic solution of the nonlinear optimization program (7) by considering the necessary and sufficient Karusch-Kuhn-Tucker conditions (10). In the setup of Theorem 4.2 these conditions simplify to (replacing $\mathbf{l}$ with $\gamma\Sigma\mathbf{l}^*$):

$$\begin{aligned} \boldsymbol{\mu} - \mathbf{x} + \boldsymbol{\Psi} &= \mathbf{0} \\ -\boldsymbol{\mu} + (\check{L}^\top - I)\boldsymbol{\Psi} + \check{L}^\top \boldsymbol{\eta} &= \mathbf{0} \\ \boldsymbol{\eta} \geqslant \mathbf{0}, \;\; \gamma L L^\top \mathbf{l}^* - L\mathbf{x} &\leqslant \mathbf{0} \\ \boldsymbol{\eta}^\top(\gamma L L^\top \mathbf{l}^* - L\mathbf{x}) &= 0 \end{aligned} \tag{21}$$

We can thus verify via direct substitution that the following

$$\tilde{\mathbf{x}} = \gamma L^\top \mathbf{l}^*, \qquad \tilde{\boldsymbol{\mu}} = \gamma(L^\top - D)\mathbf{l}^*, \qquad \tilde{\boldsymbol{\eta}} = o(\mathbf{1}) \tag{22}$$

satisfy the equations (21) asymptotically. Equations three and four in (21) are satisfied, because $\gamma L L^\top \mathbf{l}^* - L\tilde{\mathbf{x}} = \gamma L L^\top \mathbf{l}^* - L\gamma L^\top \mathbf{l}^* = \mathbf{0}$. Let us now examine equations one and two in (21). First, note that from (22)

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

and hence from the one-dimensional Mill's ratio we have

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

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

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

and for equation two (recall that $\check{L} = D^{-1}L$, so that $\check{L}^\top = L^\top D^{-1}$)

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

Thus, all of the equations in (21) are satisfied asymptotically and since (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 this pair $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\mu}})$ into $\psi(\mathbf{x}; \boldsymbol{\mu}) = \frac{\|\boldsymbol{\mu}\|^2}{2} - \mathbf{x}^\top \boldsymbol{\mu} + \sum_k \ln \overline{\Phi}(\tilde{l}_k - \mu_k)$. Using the one-dimensional Mill's ratio, $\ln \overline{\Phi}(\gamma) \simeq -\frac{1}{2}\gamma^2 - \ln \gamma - \frac{1}{2}\ln(2\pi)$, we obtain

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

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

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

In other words,

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

However, by Mill's ratio (20), we also have

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

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

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

In contrast, for the SOV estimator $\mathring{\ell}$ 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 behavior subject to the constraint $\check{L}\mathbf{x} \geqslant \gamma\check{L}L^\top \mathbf{l}^*$. This gives

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

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

estimator behaves asymptotically as

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

□

A.4 Proof of Corollary 4.1

The corollary 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 $\mathbb{P}$):

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

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

A.5 Proof of Lemma 4.1

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

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

Note that if $\alpha_i \overset{\text{def}}{=} (\ell_i-\nu_i)/\sigma_i$, $\beta_i \overset{\text{def}}{=} (u_i-\nu_i)/\sigma_i$, $p_i = \overline{\Phi}(\alpha_i)-\overline{\Phi}(\beta_i)$ and $\phi(\cdot) \equiv \phi(\cdot\,;0,1)$, then all the quantities on the left-hand side are available analytically:

$$\begin{aligned}\underline{\mathbb{E}}[X_i] &= \nu_i + \sigma_i\tfrac{\phi(\alpha_i)-\phi(\beta_i)}{p_i}\\ \mathrm{tr}(\Sigma^{-1}\underline{\mathrm{Var}}(\mathbf{X})) &= \textstyle\sum_{i=1}^d\{\Sigma^{-1}\}_{i,i}\,\sigma_i^2\Big(1+\tfrac{\alpha_i\phi(\alpha_i)-\beta_i\phi(\beta_i)}{p_i} - \Big(\tfrac{\phi(\alpha_i)-\phi(\beta_i)}{p_i}\Big)^2\Big)\\ -\underline{\mathbb{E}}[\ln\underline{\phi}(\mathbf{X})] &= \textstyle\sum_{i=1}^d\tfrac{\alpha_i\phi(\alpha_i)-\beta_i\phi(\beta_i)}{2p_i} + \ln\big(\sqrt{2\pi\exp(1)}\,\sigma_i p_i\big)\end{aligned} \tag{24}$$

Next, we establish the asymptotic behavior of $\ell_{\mathrm{L}}(\gamma)$ under the conditions of Theorem 4.2. Suppose the pair $(\tilde{\boldsymbol{\nu}},\tilde{\boldsymbol{\sigma}})$ satisfies $\mathrm{diag}^2(\tilde{\boldsymbol{\sigma}}) \simeq \Sigma$ and $\tilde{\boldsymbol{\nu}} \simeq \mathbf{l} - \gamma\mathrm{diag}^2(\tilde{\boldsymbol{\sigma}})\mathbf{l}^* = \gamma(\underline{\Sigma} - \mathrm{diag}^2(\tilde{\boldsymbol{\sigma}}))\mathbf{l}^*$ as $\gamma \uparrow \infty$. Then, $\boldsymbol{\alpha} \simeq \gamma\mathrm{diag}(\tilde{\boldsymbol{\sigma}})\mathbf{l}^*$, which in combination with $\ln\overline{\Phi}(\gamma) \simeq -\frac{1}{2}\gamma^2 - \ln(\gamma) - \frac{1}{2}\ln(2\pi)$, implies $\underline{\mathbb{E}}[\mathbf{X}] \simeq \gamma\underline{\Sigma}\mathbf{l}^*$. Hence, substituting $(\tilde{\boldsymbol{\nu}},\tilde{\boldsymbol{\sigma}})$ into

(24) and then into the left-hand-side of (23), and simplifying, we obtain

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

where the last asymptotic equivalence follows from (20). Finally, the convergence of (16) follows by applying the Pinsker-type inequality (Devroye and Györfi, 1985) in conjunction with $\sqrt{1-\exp\left(-\underline{\mathbb{E}}\ln\frac{\phi(\mathbf{X})}{f(\mathbf{X})}\right)} = \sqrt{1-\frac{1}{\ell}\exp\left(\underline{\mathbb{E}}\ln\frac{\phi(\mathbf{X};\mathbf{0},\Sigma)}{\underline{\phi}(\mathbf{X})}\right)} \leqslant \sqrt{1-\ell_{\mathrm{L}}/\ell} = o(1)$. □

References

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

Azaïs, J.-M., S. Bercu, J.-C. Fort, A. Lagnoux, and P. Lé (2010). Simultaneous confidence bands in curve prediction applied to load curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(5), 889–904.

Bolin, D. and F. Lindgren (2015). Excursion and contour uncertainty regions for latent gaussian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(1), 85–106.

Botev, Z. I., P. L'Ecuyer, and B. Tuffin (2013). Markov chain importance sampling with applications to rare event probability estimation. *Statistics and Computing 23*(2), 271–285.

Botts, C. (2013). An accept-reject algorithm for the positive multivariate normal distribution. *Computational Statistics 28*(4), 1749–1773.

Chopin, N. (2011). Fast simulation of truncated Gaussian distributions. *Statistics and Computing 21*(2), 275–288.

Craig, P. (2008). A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 227–243.

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

Devroye, L. and L. Györfi (1985). *Nonparametric density estimation: the L1 view*, Volume 119. John Wiley & Sons Inc.

Fernández, P. J., P. A. Ferrari, and S. P. Grynberg (2007). Perfectly random sampling of truncated multinormal distributions. *Advances in Applied Probability 39*(4), 973–990.

Galton, F. (1889). *Natural inheritance*, Volume 42. Macmillan.

Gassmann, H., I. Deák, and T. Szántai (2002). Computing multivariate normal probabilities: A new look. *Journal of Computational and Graphical Statistics 11*(4), 920–949.

Gassmann, H. I. (2003). Multivariate normal probabilities: implementing an old idea of Plackett's. *Journal of Computational and Graphical Statistics 12*(3), 731–752.

Genton, M. G., Y. Ma, and H. Sang (2011). On the likelihood function of Gaussian max-stable processes. *Biometrika 98*(2), 481–488.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics 1*(2), 141–149.

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. Springer.

Gerber, M. and N. Chopin (2015). Sequential Quasi-Monte-Carlo sampling. *J. R. Statist. Soc. B 77*(3), 1–44.

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: Proceedings of the 23rd symposium on the interface*, pp. 571–578. Citeseer.

Grün, B. and K. Hornik (2012). Modelling human immunodeficiency virus ribonucleic acid levels with finite mixtures for censored longitudinal data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(2), 201–218.

Hajivassiliou, V. A. and D. L. McFadden (1998). The method of simulated scores for the estimation of LDV models. *Econometrica 66*(4), 863–896.

Hashorva, E. and J. Hüsler (2003). On multivariate gaussian tails. *Annals of the Institute of Statistical Mathematics 55*(3), 507–522.

Hayter, A. J. and Y. Lin (2012). The evaluation of two-sided orthant probabilities for a quadrivariate normal distribution. *Computational Statistics 27*(3), 459–471.

Hayter, A. J. and Y. Lin (2013). The evaluation of trivariate normal probabilities defined by linear inequalities. *Journal of Statistical Computation and Simulation 83*(4), 668–676.

Huser, R. and A. C. Davison (2013). Composite likelihood estimation for the Brown–Resnick process. *Biometrika 100*(2), 511–518.

Joe, H. (1995). Approximations to multivariate normal rectangle probabilities based on conditional expectations. *Journal of the American Statistical Association 90*(431), 957–964.

Koop, G., D. J. Poirier, and J. L. Tobias (2007). *Bayesian econometric methods*, Volume 7. Cambridge University Press.

Kroese, D. P., T. Taimre, and Z. I. Botev (2011). *Handbook of Monte Carlo Methods*, Volume 706. 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 (TOMACS) 20*(1), 6.

Miwa, T., A. J. Hayter, and S. Kuriki (2003). The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 223–234.

Nomura, N. (2014a). Computation of multivariate normal probabilities with polar coordinate systems. *Journal of Statistical Computation and Simulation 84*(3), 491–512.

Nomura, N. (2014b). Evaluation of Gaussian orthant probabilities based on orthogonal projections to subspaces. *Statistics and Computing*, in press.

Philippe, A. and C. P. Robert (2003). Perfect simulation of positive Gaussian distributions. *Statistics and Computing 13*(2), 179–186.

Powell, M. J. D. (1970). A hybrid method for nonlinear equations. *Numerical methods for nonlinear algebraic equations 7*, 87–114.

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

Sándor, Z. and P. András (2004). Alternative sampling methods for estimating multivariate normal probabilities. *Journal of Econometrics 120*(2), 207–234.

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

Tuffin, B. (1999). Bounded normal approximation in simulations of highly reliable markovian systems. *Journal of Applied Probability 36*(4), 974–986.

Vijverberg, W. (1997). Monte Carlo evaluation of multivariate normal probabilities. *Journal of Econometrics 76*(1), 281–307.

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