Perfect simulation of positive Gaussian distributions†

Anne Philippe

Laboratoire de Statistique et Probabilités, FRE CNRS 2222, Université de Lille 1, Bat M2 59655 Villeneuve D'Ascq cedex, France

Christian P. Robert

CREST, INSEE, and CEREMADE, Université Paris Dauphine, 75775 Paris cedex 16, France

Summary. We provide an exact simulation algorithm that produces variables from truncated Gaussian distributions on $(\mathbb{R}_+)^p$ via a perfect sampling scheme, based on stochastic ordering and slice sampling, since accept–reject algorithms like those of Geweke (1991) or Robert (1994) are difficult to extend to higher dimensions.

Keywords: Accept reject, coupling from the past, slice sampling, stochastic monotonicity.

1. Introduction

The simulation from a normal distribution $\mathcal{N}_p(\mu, \Sigma)$, restricted to a subset of $\mathbb{R}^p$, and in particular, to the positive quadrant $(\mathbb{R}_+)^p$, with density proportional to

$$\varphi(x) = \exp\left(-\frac{1}{2}(x-\mu)^t\,\Sigma^{-1}\,(x-\mu)\right)\mathbb{I}_{(\mathbb{R}_+)^p}(x)\,, \tag{1.1}$$

called abusively "positive Gaussian distribution", is a recurrent problem in many areas, including signal processing (clipping) and econometrics (factor models, multivariate tobit models). See, e.g., Hajivassiliou, McFadden and Ruud (1996) for examples.

There have been proposals in the past to simulate these truncated distributions by using Gibbs sampling based on exact sampling algorithms in dimension one, since the conditional distributions are all truncated normal distributions. See, for instance, Geweke (1991) or Robert (1994), for exact (accept-reject) sampling algorithms in dimension one and details about the Gibbs algorithms.

These solutions are, however, unsatisfactory in that they do not produce exact simulations from the multivariate truncated normal distribution, except in an aymptotic sense practitioners are unlikely to accept. Standard simulation methods such as accept-reject methods are also difficult to implement since the performance of a proposal distribution is bound to depend on the shape of the normal distribution on the positive quadrant. For instance, the simplest proposal, based on simulating from the normal distribution $\mathcal{N}_p(\mu, \Sigma)$ until all components are positive, offers no lower bound on the probability of acceptance.

We show in Section 2 that a perfect sampling algorithm, based on coupling from the past (CFTP) and slice sampling, is available to simulate directly and exactly from the multivariate truncated normal distribution, describe the algorithm in Section 3 and illustrate this fact in dimension two in Section 4. Appendix A details the construction of the algorithm in dimension two.

†Work partially supported by EU TMR network ERB–FMRX–CT96–0095 on *'Computational and Statistical Methods for the Analysis of Spatial Data'*.

2. Perfect sampling

The general theory of perfect (or exact) sampling, introduced by Propp and Wilson (1996), is now comprehensively covered in surveys like Dimakos (2001) or Casella, Lavine and Robert (2001), and we refer the reader to these papers for an introduction to the topic. In this paper, we make use more particularly of the specific type of perfect sampling recently developped in Mira, Møller and Roberts (2001), based on coupling from the past (CFTP), slice sampling and stochastic ordering of Markov chains. We simply recall here that CFTP consists in running an MCMC sampler with parallel Markov chains using all possible starting points at time $-T$, with T "large enough" for all chains to be identical by time 0. While this event, called *coalescence*, has probability 0 to occur for independent parallel chains, in most setups, it can gain a positive probability when the chains are *coupled*, that is, based on the same sequence of uniform variates.

We also refer the reader to another source, namely Robert and Casella (1999), for a general description of slice sampling, which is at its core a Gibbs sampling implementation of the simulation from the joint distribution

$$(X, U) \sim \mathcal{U}_{\{(x,u);\ 0 \le u \le f(x)\}}\,,$$

which enjoys f as its marginal distribution (in x). (See also Damien, Wakefield and Walker, 1999.)

An appealing feature of slice samplers in the setting of perfect sampling, pointed out by Mira *et al.* (2001), is that there exists a natural stochastic ordering, which is induced by $\varphi(\cdot)$. Hence, monotonicity arguments can be invoked to reduce the number of parallel chains to two parallel chains only. (See also Casella *et al.*, 2001, for a similar derivation of a perfect slice sampler in the setup of mixtures of distributions.)

More precisely, note first that, if $\varphi(x_1) \le \varphi(x_2)$, the corresponding slices satisfy

$$\mathcal{A}_2 = \{x;\ \varphi(x) \ge u\,\varphi(x_2)\} \subset \mathcal{A}_1 = \{x;\ \varphi(x) \ge u\,\varphi(x_1)\}.$$

Therefore, simulation from a uniform distribution on $\mathcal{A}_2$ can first proceed by acceptance–rejection of a uniform sampling on $\mathcal{A}_1$. In other words, and from a perfect sampling point of view, this property induces a natural possibility of coalescence: if $x_1' \sim \mathcal{U}(\mathcal{A}_1)$ also belongs to $\mathcal{A}_2$, this realization is acceptable as a simulation from $\mathcal{U}(\mathcal{A}_2)$ and both chains coalesce, that is, they are identical from this epoch; if x_1' does not belong to $\mathcal{A}_2$, the value x_2' simulated subsequently from the uniform distribution on $\mathcal{A}_2$, preserves the ordering $\varphi(x_1') \le \varphi(x_2')$. Therefore, exploiting this possibility for coalescence at each epoch ensures that two Markov chains such that

$$\varphi(x_1^{(0)}) \le \varphi(x_2^{(0)})$$

at time 0 remain ordered in the sense that

$$\varphi(x_1^{(t)}) \le \varphi(x_2^{(t)})$$

for every t, the inequality turning into an equality for t large enough.

Secondly, it happens that, in the case of truncated normal distributions, there exist both a maximal and a minimal element in $(\mathbb{R}_+)^p$, $\tilde{1}$ and $\tilde{0}$, for the order induced by φ, since this density is bounded. Moreover, these maximal and minimal elements can be identified since $\tilde{1}$ is either μ, if $\mu \in (\mathbb{R}_+)^p$, or a point on the boundary of $(\mathbb{R}_+)^p$ otherwise (see Theorem 1 in the Appendix for an example in dimension 2), while $\tilde{0}$ is ∞. Thus, monotone CFTP in the sense of Propp and Wilson (1996) applies: it is sufficient to run two chains starting from $\tilde{1}$ and $\tilde{0}$, respectively, at time $-T$, and increase T, that is, go back further in time, till both chains have coalesced by time 0. By virtue of a straightforward monotonicity argument, the collection of all chains in between the two extreme chains, even when this collection is uncountable, have coalesced by the time those two coalesce.

3. The exact simulation algorithm

The perfect sampling (CFTP) algorithm looks as follows:

$T \leftarrow -1$
Repeat

Take $t \leftarrow T$, $\omega_0^{(T)} \leftarrow \tilde{0}$ and $\omega_1^{(T)} \leftarrow \tilde{1}$

While $t < 0$, do

0. If $t < T/2$, generate $seed^{(t)}, u^{(t)} \sim \mathcal{U}([0,1])$
1. $\epsilon \leftarrow u_1^{(t)} \varphi(\omega_0^{(t)})$
2. Generate $\omega_0^{(t+1)}$ from

$$\mathcal{U}\left(\left\{x \in (\mathbb{R}_+)^p;\ (x-\mu)^t\, \Sigma^{-1}\, (x-\mu) \leq -2\log(\epsilon)\right\}\right) \tag{3.1}$$

starting with $seed^{(t)}$.

3. If $\varphi(\omega_0^{(t+1)}) \geq u_1^{(t)} \varphi(\omega_1^{(t)}) = \epsilon'$, take $\omega_1^{(t+1)} \leftarrow \omega_0^{(t+1)}$
Otherwise

Generate $\omega_1^{(t+1)}$ from

$$\mathcal{U}\left(\left\{x \in (\mathbb{R}_+)^p;\ (x-\mu)^t\, \Sigma^{-1}\, (x-\mu) \leq -2\log(\epsilon')\right\}\right)$$

starting with $seed^{(t)}$.

4. $t \leftarrow t+1$

$not.coalescence \leftarrow \{\omega_0^{(0)} \neq \omega_1^{(0)}\}$

$T \leftarrow 2 * T$

while $not.coalescence$

This pseudo-code requires some detailed explanations. First, the introduction of the variables $seed^{(t)}$ and $u^{(t)}$ at step 0. and the use of $seed^{(t)}$ in steps 2. and 3. ensures that the coupling from the past algorithm is correct, that is, that it produces a simulation from the distribution of interest φ. In fact, as described in Propp and Wilson (1996), the validity of the algorithm requires the *same* sequence of random variables (typically, those are uniform random variables) to be used each time the epoch t ($t = 0, -1, \ldots$) is visited. In other words, one must determine a single realisation of a sequence of random transforms $(\Psi_t)_t$ such that

$$\omega^{(t+1)} = \Psi_t\left(\omega^{(t)}\right)$$

corresponds to the transition from $\omega^{(t)}$ to $\omega^{(t+1)}$, whatever the value of $\omega^{(t)}$ is. Since simulation from (3.1) can only be done via an accept–reject algorithm (described below), the number of uniforms used at each epoch varies, but the sequence of uniforms used by this accept–reject algorithm will always be the same if we start the uniform generator on the computer with the same seed $seed^{(t)}$.

Second, simulation from (3.1) requires accept–reject steps, based on a uniform proposal on an hypercube (or *box*) $\mathcal{B}$ that contains the ellipsoid

$$\mathcal{E} = \left\{x \in (\mathbb{R}_+)^p;\ (x-\mu)^t\, \Sigma^{-1}\,(x-\mu) \le -2\log(\epsilon)\right\}\ ,$$

and on which uniform simulation is possible. The specific construction of the box $\mathcal{B}$ is proposed for $p=2$ in the appendix A. In general, the box $\mathcal{B}$ is determined by the extreme points of $\mathcal{E}$ in every direction and $seed^{(t)}$ is a uniform random variable on $[0,1]^p$. (Note that, in the very special case when $\mu = \mathbf{0}$ and $\Sigma = \sigma^2 I_p$, the finite sequence of uniforms is exactly the same.)

Third, there is a difficulty with the extreme point $\tilde{0}$. While $\tilde{0} = \infty$ does exist, it is not possible to simulate uniformly on the slice $\{\varphi(x) \ge 0\}$ because it is equal to $(\mathbb{R}_+)^p$. We thus replace ∞ with a large enough value of $\tilde{0}$, that is, such that $\varphi(\tilde{0}) \ll \varphi(\tilde{1})$ to implement the method.

4. Some illustrations

In this section, we consider two examples of simulations of two dimensional truncated Gaussian distributions, namely

Example 1 $\mu = (2,2)$ and $\Sigma = \begin{pmatrix} 2.5 & -2.5 \\ -2.5 & 3.5 \end{pmatrix}$

Example 2 $\mu = (1,-5)$ and $\Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 2 \end{pmatrix}$

Note that the marginal densities are not truncated Gaussian anymore. For instance,

$$\varphi_1(x_1) \propto e^{-\frac{1}{2\sigma_{11}}(x_1-\mu_1)^2}\left(1-\Phi\left(-\frac{\sqrt{\sigma_{11}}}{\sqrt{|\Sigma|}}\left(\mu_2+\frac{\sigma_{12}}{\sigma_{11}}(x_1-\mu_1)\right)\right)\right)\mathbb{I}_{\mathbb{R}^+}(x_1)$$

where Φ is the standard Gaussian cdf and the covariance matrix is given by

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\ .$$

Appendix A provides the details about the calibration of the accept–reject algorithm and the construction of the “boxes” that contain the slices to simulate from.

Figure 1 shows a typical sequence of trials necessary to achieve coalescence. (The moves of both the upper and lower chains are different on each graph, that is, for each starting time T, since the requirement on the uniforms is to use the same starting seed at a given time t.) As one can see from the graphs, the chain starting at $\tilde{0}$ (chosen as $(10,10)$ in this case) requires several steps to reach the region of interest, slow moves in regions of small posterior densities being characteristic of the slice sampler.

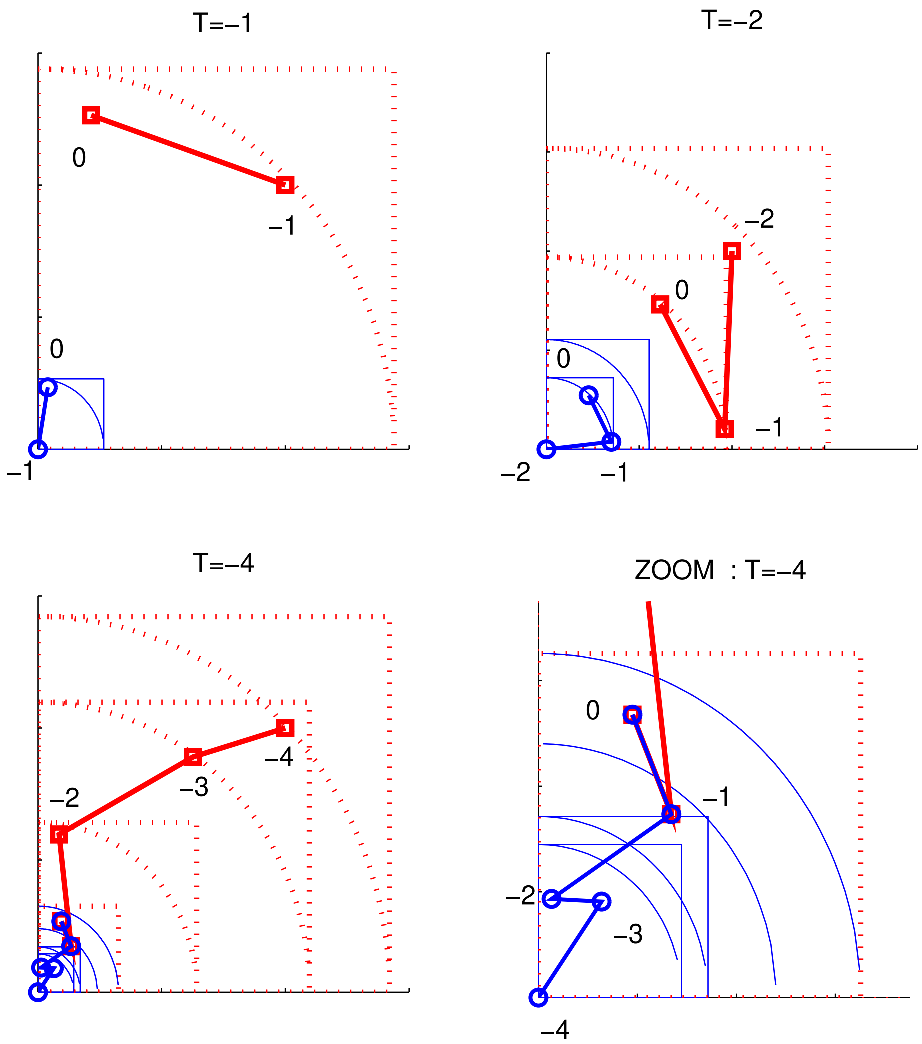

Fig. 1. A coalescence story for $\mu = (0,0)^t$ and $\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$. The lower right graph represents a subset of the lower left graph. It indicates that coalescence actually took place at time $t = -1$. (The numbers on the graphs represent the time at which the corresponding points were simulated.)

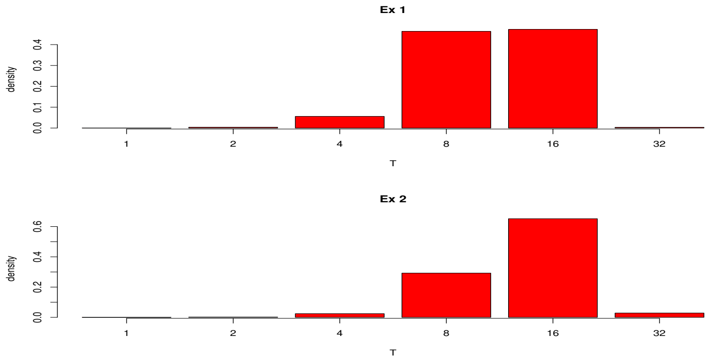

Fig. 2. Histogram of the coalescence time for both examples (obtained from $10,000$ independent replications and $\tilde{0} = (100, 100)$)

As mentioned above, this graph is typical of the coalescence process. Figure 2 shows how the coalescence times are distributed across $10,000$ experiments. In most cases, coalescence occurs between $T = -8$ and $T = -16$ horizons. (Starting with a larger value of $\tilde{0}$ increases the coalescence times, without modifying the shape of the sample histograms.)

Figures 3 and 4 provide a dot representation of the sample produced by our algorithm for two sets of parameters μ and Σ, along with a adequation between the marginal histograms of these samples and the corresponding true marginal densities.

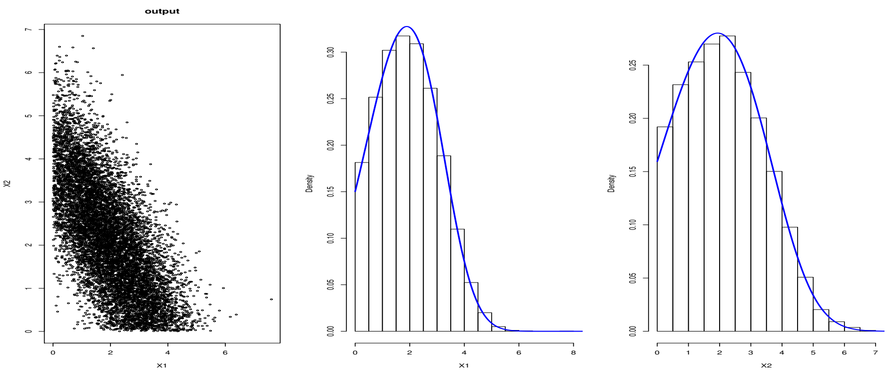

Fig. 3. Output and marginal histograms produced by the perfect sampler for Example 1 with $\tilde{0} = (100, 100)$. (The sample size is $10,000$.)

References

Casella, G., Lavine, M. and Robert, C.P. (2001) Explaining the perfect sampler. *The American Statistician* (to appear).

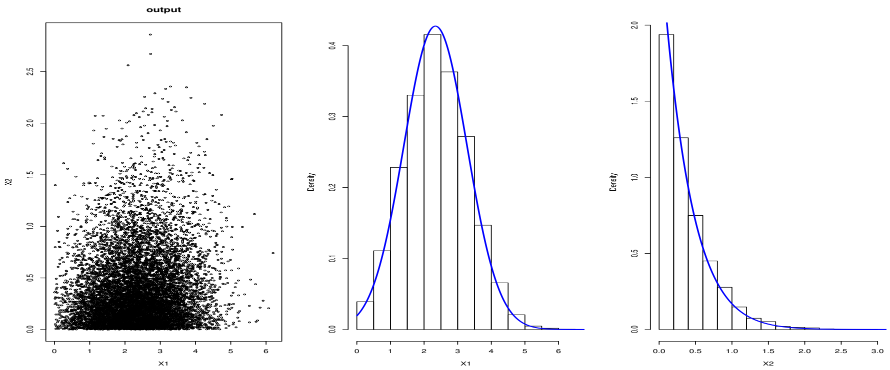

Fig. 4. Output and marginal histograms produced by the perfect sampler for Example 2 with $\tilde{0}_{(}100, 100)$. (The sample size is $10,000$.)

Casella, G., Mengersen, K.L., Robert, C.P. and Titterington, D.M. (2001) Perfect samplers for mixtures of distributions. *J. Royal Statist. Soc.* (Ser. B) (to appear).

Damien, P., Wakefield, J. and Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Royal Statist. Soc.* (Ser. B) **61**, 331–344.

Dimakos, X. (2001) A guide to exact simulation. *Intern. Statist. Rev.* **69**(1), 27–48.

Geweke, J. (1991) Efficient simulation from the multivariate normal and Student t-distributions subject to linear constraints. *Computer Sciences and Statistics: Proc. 23d Symp. Interface*, 571–577.

Hajivassiliou, V.A., McFadden, D. and Ruud, P.A. (1996) Simulation of multivariate normal rectangle probabilities and their derivatives: theoretical and computational results. *J. Econometrics* **72**, 85–134.

Hobert, J.P., Robert, C.P. and Titterington, D.M. (1999) On perfect simulation for some mixtures of distributions. *Statistics and Computing* **9**, 287–298.

Mira, A., Møller, J. and Roberts, G.O. (2001) Perfect slice samplers. *J. Royal Statist. Soc.* (Ser. B) **63**(3), 593–606.

Propp, J. G. and Wilson, D. B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9** 223-252.

Robert, C.P. (1995) Simulation of truncated normal variables. *Statistics and Computing* **5**, 121–125.

Robert, C.P. and Casella, G. (1999) *Monte Carlo Statistical Methods.* Springer Verlag, New York.

Wilson, D.B. (1998) Annotated bibliography of perfectly random sampling with Markov chains. In D. Aldous and J. Propp, editors, *Microsurveys in Discrete Probability*, volume 41 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 209–220. American Mathematical Society.

A. The special case $p = 2$

In this appendix, we consider more specifically the case $p = 2$, that is, a normal $\mathcal{N}_2(\mu, \Sigma)$ distribution, restricted to $\mathbb{R}_+^2$, with

$$\mu = (\mu_1, \mu_2)^t, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

We denote by Λ the diagonal matrix of eigenvalues $\{\lambda_1, \lambda_2\}$ of Σ and Q the matrix of the corresponding eigenvectors $\{Q_1, Q_2\}$. The matrix Σ can be expressed as $\Sigma = Q^t \Lambda Q$ and its inverse $\Sigma^{-1} = Q\Lambda^{-1}Q^t$, where M^t denotes the matrix transpose of M.

The argument of the maximum of φ on $(\mathbb{R}_+)^2$ can then be obtained as follows:

THEOREM 1. *The density φ is bounded and its maximum x^* in $(\mathbb{R}_+)^2$ is given by*

$$x^* = \begin{cases} \mu & \text{if } \mu \in \mathbb{R}_+^2, \\ x^*_{\max} & \text{otherwise.} \end{cases}$$

where

$$x^*_{\max} = \arg\max_{x \in \mathcal{M}^+} (x - \mu)^t \Sigma^{-1} (x - \mu),$$

with

$$\mathcal{M}^+ = \left\{ (0,0)\,,\ (\mu_1 + \mu_2 \frac{\sigma_{12}}{\sigma_{22}}, 0)\,,\ (0, \mu_2 + \mu_1 \frac{\sigma_{12}}{\sigma_{11}}) \right\} \bigcap (\mathbb{R}_+)^2.$$

We now detail the simulation of random variables from the uniform distribution on $\{\varphi(x) > \epsilon_t \varphi(x_{t-1})\}$. First, we can rewrite this set as

$$\begin{aligned} \{x : \varphi(x) > \epsilon_t \varphi(x_{t-1})\} &= \{x : (x-\mu)^t \Sigma^{-1} (x - \mu) < K_t,\ x \in (\mathbb{R}_+)^2\} \\ &= \{y : y^t \Lambda^{-1} y < K_t,\ y = Q^t(x - \mu),\ \ x \in (\mathbb{R}_+)^2\} \end{aligned}$$

with $K_t = -2\log(\epsilon_t) + (x_{t-1} - \mu)^t \Sigma^{-1} (x_{t-1} - \mu)$

We denote $(x_1, x_2)_o$ the coordinates in (O, i, j) and $(y_1, y_2)_e$ the coordinates in (μ, Q_1, Q_2). We have

$$(y_1, y_2)_e^t = Q^t((x_1, x_2)_o^t - \mu) \text{ and } (x_1, x_2)_o^t = Q(y_1, y_2)_e^t + \mu\,.$$

Let $\mathcal{B}_U = [0, 1]^2$. We construct a box of the form

$$\mathcal{B}_t = \{(h_1\, u_1 + t_1, h_2\, u_2 + t_2)_e, (u_1, u_2)_e \in \mathcal{B}_U\}$$

such that

$$\{x : \varphi(x) > \epsilon_t \varphi(x_{t-1})\} \subseteq \mathcal{B}_t\,.$$

In order to determine the vertices of the box $\mathcal{B}_t$, consider the set

$$\mathcal{S} = \{(\sqrt{\lambda_1 K}, 0)_e, (-\sqrt{\lambda_1 K}, 0)_e, (0, \sqrt{\lambda_2 K})_e, (0, -\sqrt{\lambda_2 K})_e\}$$

and $\mathcal{S}^+ = \mathcal{S} \bigcap (\mathbb{R}_+)^2$

	Example 1	Example 2
Mean	1.64	2.32
Variance	1.24	4.68

Table 1. Number of uniforms simulated by the accept-reject algorithm.

LEMMA 1. *If* $|\mathcal{S}^+| = 4$ *then*

$$(h_1, h_2, t_1, t_2) = (2\sqrt{\lambda_1 K}, 2\sqrt{\lambda_2 K}, \min\{0, \sqrt{\lambda_1 K}, -\sqrt{\lambda_1 K}\}, \min\{0, \sqrt{\lambda_2 K}, -\sqrt{\lambda_2 K}\})$$

When $|\mathcal{S}^+| < 4$, we calculate the points $(a^{(i)})$, $i \in I_1$ where the $(a^{(i)}, 0)_o$'s are the roots of the equation

$$(x-\mu)^t \Sigma^{-1} (x-\mu) = K_t$$

on the horizontal axis, and the points $b^{(i)}$, $i \in I_2$ where the $(0, b^{(i)})_o$'s are the roots of the equation

$$(x-\mu)^t \Sigma^{-1} (x-\mu) = K_t$$

on the vertical axis.

We then construct the set of points

$$\begin{aligned} \mathcal{W} &= \begin{cases} \mathcal{S}^+ \bigcup \{Q^t((a^{(i)},0)^t - \mu), i \in I_1\} \bigcup \{Q^t((0,b^{(i)})^t - \mu), i \in I_2\} \bigcup \{-Q^t\mu\} & \text{if } |I_1| = |I_2| = 1, \\ \mathcal{S}^+ \bigcup \{Q^t((a^{(i)},0)^t - \mu), i \in I_1\} \bigcup \{Q^t((0,b^{(i)})^t - \mu), i \in I_2\} & \text{otherwise} \end{cases} \\ &= \left\{ (y_1^{(i)}, y_2^{(i)})_e, i \in I \right\} \end{aligned}$$

and deduce the vertices of $\mathcal{B}_t$ as follows:

LEMMA 2. *If* $|\mathcal{S}^+| < 4$ *then*

$$\begin{aligned} h_1 &= \max\left\{y_1^{(i)},\ i \in I\right\} - \min\left\{y_1^{(i)},\ i \in I\right\} \\ h_2 &= \max\left\{y_2^{(i)},\ i \in I\right\} - \min\left\{y_2^{(i)},\ i \in I\right\} \\ t_1 &= \min\left\{y_1^{(i)},\ i \in I\right\} \\ t_2 &= \min\left\{y_2^{(i)},\ i \in I\right\} \end{aligned}$$

Once the box $\mathcal{B}_t$ is constructed, a straightforward simulation from the uniform distribution on the ellipsoid

$$\{x : \varphi(x) > \epsilon_t \varphi(x_{t-1})\}$$

is then to simulate uniformly on $\mathcal{B}_t$ until the simulation belongs to the ellipsoid. Note that, since the box $\mathcal{B}_t$ is an affine transformation of $\mathcal{D} = [0,1] \times [0,1]$, it is sufficient to simulate a sequence of uniforms on $\mathcal{D}$. Table 1 gives the performance of this accept-reject algorithm for both examples treated in Figures 3 and 4.