Computational Statistics (2004) 19:301–315

Automated Rejection Sampling from Product of Distributions

G. Marrelec and H. Benali

INSERM U494, CHU Pitié-Salpêtrière
91, boulevard de l'Hôpital
F-75634 Paris

Summary

Sampling from probability density functions (pdfs) has become more and more important in many areas of applied science, and has therefore been the subject of great attention. Many sampling procedures proposed allow for approximate or asymptotic sampling. On the other hand, very few methods allow for exact sampling. Direct sampling of standard pdfs is feasible, but sampling of much more complicated pdfs is often required. Rejection sampling allows to exactly sample from univariate pdfs, but has the huge drawback of needing a case-by-case calculation of a comparison function that often reveals as a tremendous chore, whose results dramatically affect the efficiency of the sampling procedure. In this paper, we restrict ourselves to a pdf that is proportional to a product of standard distributions. From there, we show that an automated selection of both the comparison function and the upper bound is possible. Moreover, this choice is performed in order to optimize the sampling efficiency among a range of potential solutions. Finally, the method is illustrated on a few examples.

Keywords: Rejection sampling, Exact sampling, Gibbs sampling, MCMC methods

1 Introduction

Sampling from probability density functions (pdfs) has become a crucial topic in many areas of applied science. There is a need for sampling whenever statistical methods are developed, may it be for approximate estimation or model diagnosis (Gelman et al. 1998). Numerous schemes have consequently been devised in order to sample from a given pdf, ranging from approximate sampling to asymptotic sampling to exact sampling.

Approximate sampling provides samples that are only approximately related to the pdf that needs sampling. The most used is possibly importance sampling (Ruanaidh and Fitzgerald 1996), that samples from an approximate pdf and gives a weight to every sample to compensate for the approximation. In this case, the reliability only depends on the relation between the pdf to sample and the approximation chosen. Unfortunately, the error between an approximate sample and one originating from the true pdf can neither be estimated nor reduced by increased computational power. Approximate sampling is therefore shun and its use restricted to cases where no other method applies.

Asymptotic methods asymptotically produce samples from a given pdf. The most common such method is Markov Chain Monte Carlo (MCMC) and Gibbs sampling as a special case (Gelman et al. 1998). Both can be used to sample from many complicated pdfs encountered in statistics. However, they may take a long time to converge towards a true sample, and convergence diagnosis is still a topic of ongoing research.

Finally, *exact sampling* includes methods that can provide samples that are exactly generated by the pdf under study. A large amount of literature can be found on generating exact samples from standard pdfs (Devroye 1986). Such methods are usually designed for a particular distribution (*e.g.* Gaussian, Poisson, Cauchy). In many cases, though, standard pdfs are not sufficient, and several efficient methods have been proposed to deal with this issue. In the last years, some papers have appeared on automatic (also called "black-box") methods (Devroye 1986). The methods proposed can sample from a large family of pdfs as long as some information (*e.g.* mode of the pdf) is available.

Rejection sampling is a conceptually very simple method to perform exact sampling from univariate pdfs (Press et al. 1992; Tanner 1994). It is a three-step scheme that consists of majoring the pdf of interest $p(x)$ by the product of another pdf $q(x)$– the *comparison function*–by a constant c–the rejection constant–:

$$p(x) \leq c \cdot q(x), \tag{1}$$

then sampling from the comparison function $q(x)$, and finally deciding whether the point sampled should be kept or rejected based on the acceptance ratio $p(x)/cq(x)$. The closer the majoring function $cq(x)$ is to the true pdf $p(x)$, the more samples are accepted, and therefore the more efficient the algorithm is. The preliminary choice of the comparison function and of the upper bound defines the rejection rate, which, in turn, dramatically affects the sampling in terms of time needed to obtain a sample

of given size. The scientist willing to apply rejection sampling has first to undergo a long analysis before the sampling procedure can be applied. Therefore, construction plans have to be developed in order to speed up this first step.

The ratio-of-uniform method introduced by Kinderman and Monahan (1977) is one of such methods. This technique can be used to sample from a large variety of pdfs whose densities are proportional to some known functions. This is performed by sampling uniformly from a n-dimensional region that lies under the plot of the given pdf. It has become a popular method to generate samples, since it results in an exact, fast and efficient algorithm. However, it seems to be difficult for most pdfs to obtain the necessary rectangle enclosing the region of acceptance for the multivariate extension of the ratio-of-uniform method (Leydold 1998).

The Adaptive Rejection Sampling (ARS) method by Gilks and Wild (1992) usually assists the routine use of MCMC sampling methods. This technique can be used to efficiently sample from any univariate distribution whose density function is log-concave. To automatically generate a majoring function, we only need to provide the pdf mode. ARS then proceeds by constructing an envelope function of the log of the pdf. However, log-concavity does not hold for all pdfs. Adaptive Rejection Metropolis Sampling (ARMS) deals with this case by performing a Metropolis step on each point accepted at an ARS rejection step. These algorithms can generate samples from a large family of pdfs as long as the mode of the pdf is available.

Several approaches for the generation of samples from multivariate distributions also exist, such as the decomposition and rejection method by Dagpunar (1988). The majoring function suggested for the multivariate rejection step is the product of the marginal densities. However, this method still requires the choice of the rejection constant.

Leydold and Hormann have recently developed an algorithm for log-concave multivariate distributions (Leydold 1998). This algorithm uses the idea of the transformed density rejection which is presented in Gilks and Wild (1992). Although this algorithm works, it is very slow, since the construction of the majoring function uses points on each side of the mode of the multivariate pdf which is decomposed in polyhedra (Leydold 1998).

A recent algorithm, the so-called "Slice Sampling" method, has been proposed by Neal (2000) for multivariate pdfs. This method samples from a pdf by uniformly sampling from the region under the plot of the pdf, and then by looking only on the "horizontal slice" defined by the current vertical position. Again, the same problems arise as in ratio-of-uniform sampling, amplified by the effect of dimensionality.

In this paper, we place ourselves in the particular case where the (possibly multivariate and/or multimodal) pdf that needs sampling is proportional to the product of pdfs for which exact sampling procedures are available. We show that, in this particular case, rejection sampling can be performed, with a very interesting feature: automated selection of both the comparison function and the upper bound can be achieved. Moreover, these choices can be done so that they maximize the probability of acceptance among a certain range of potential functions. We call this method Automated Rejection Sampling from Product of distributions (ARSP).

Applications of this model are numerous. They include Gibbs sampling, since it is often the case that the conditional pdf that needs sampling at every step has the form assumed in this article. More generally, MCMC methods are also concerned, if an approximate form of the pdf as a product of laws is available. Markov Random Fields (MRF) are also aimed at, since in many cases the conditional pdf can be formulated as a product of exponentials according to Clifford theorem (Winkler 1995). Finally, in Bayesian analysis, the posterior pdf is proportional to the product of the likelihood by a prior according to Bayes' theorem. If one can independently sample from both pdfs, then the schemed proposed here can be applied and it is possible to sample from the posterior pdf. Furthermore, if the data are assumed to be independent samples from a distribution, the likelihood can be expressed as the product of several pdfs, and the method proposed here can again be applied.

The outcome of this article is as follows: In Section 2, we set the model under study (Section 2.2) and propose a method to sample from it (Section 2.3), where both the comparison function and the rejection constant are automatically selected in order to maximize the acceptance rate. Section 3 then illustrates the features of the method with various simulations, regarding the number of distributions in the product and the dimensionality of the variate.

2 Method

2.1 Notation

In the following, x denotes a real number, $\boldsymbol{x}$ a vector, and $\boldsymbol{X}$ a matrix. $\mathbb{R}$ is the set of all real numbers."^{t}" is the regular matrix transposition. "$\equiv$" relates two expressions that are set equal by definition, and "$\propto$" two expressions that are proportional. "$\sim$" defines a random variable. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the Gaussian distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\mathcal{B}(r)$ for the discrete Bernoulli distribution with parameter r. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}_0)$ represents the value of the Gaussian function with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ evaluated at point $\boldsymbol{x}_0$. "$\widetilde{\boldsymbol{x}} \cong \mathcal{L}(\boldsymbol{\theta})$" means that the random variable $\widetilde{\boldsymbol{x}}$ is $\mathcal{L}$-distributed with parameter vector $\boldsymbol{\theta}$.

2.2 The issue

It is henceforth assumed that the pdf $p(\boldsymbol{x})$ that needs sampling is proportional to a product of standard pdfs $f_n(\boldsymbol{x})$, $1 \leq n \leq N$:

$$p(\boldsymbol{x}) = k \prod_{n=1}^{N} f_n(\boldsymbol{x}). \tag{2}$$

By "standard" is meant that a sampling procedure is available and that there is an easy analytical or numerical access (*e.g.* through the sampling procedure) to various parameters (principally the mode, but also the mean and the variance if possible). The sampling problem consists of finding a series $(\boldsymbol{x}^{[1]}, \boldsymbol{x}^{[2]}, \ldots, \boldsymbol{x}^{[J]})$ of vectors

that are *exactly* distributed according to $p(\boldsymbol{x})$. This is achieved through rejection sampling.

2.3 Rejection sampling

In order to perform rejection sampling, a comparison function and the corresponding rejection constant are needed. Assuming that the model given in Equation (2) holds, it is possible to propose a whole family of functions that qualify as comparison functions.

Theorem 1 *Let* $\alpha_n \equiv \prod_{m \neq n} \sup_{\boldsymbol{x}} [f_m(\boldsymbol{x})]$, $n = 1, \ldots, N$, *and* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)^t$ *a vector whose components are positive and add up to 1. Then the following inequality holds:*

$$p(\boldsymbol{x}) \leq c_{\boldsymbol{\lambda}} q_{\boldsymbol{\lambda}}(\boldsymbol{x}), \tag{3}$$

where

$$c_{\boldsymbol{\lambda}} \equiv k \sum_{n=1}^{N} \lambda_n \alpha_n \qquad \text{and} \qquad q_{\boldsymbol{\lambda}}(\boldsymbol{x}) \equiv \frac{1}{\sum_{n=1}^{N} \lambda_n \alpha_n} \sum_{n=1}^{N} \lambda_n \alpha_n f_n(\boldsymbol{x}). \tag{4}$$

Proof: For every n, it is possible to major $p(\boldsymbol{x})$ by $k\alpha_n f_n(\boldsymbol{x})$, and therefore by a weighted average of these factors:

$$p(\boldsymbol{x}) \leq k \sum_{n=1}^{N} \lambda_n \alpha_n f_n(\boldsymbol{x}),$$

where $(\lambda_1, \ldots, \lambda_N)^t$ is defined as in the theorem. Explicit consideration of the normalization factor in the majoring term directly leads to Equations (3) and (4). □

The function $q_{\boldsymbol{\lambda}}(\boldsymbol{x})$ being normalized, it can be considered as a pdf. Assuming further that sampling from this function is feasible, consideration of Equations (1) and (3) indicates that $q_{\boldsymbol{\lambda}}(\boldsymbol{x})$ can be considered as a comparison function and $c_{\boldsymbol{\lambda}}$ as the corresponding rejection constant for rejection sampling. In this case, the sampling procedure reads:

1. sample $\boldsymbol{x}^{[j]}$ according to $\widetilde{\boldsymbol{x}} \cong q_{\boldsymbol{\lambda}}(\boldsymbol{x})$;
2. sample $u^{[j]}$ according to $(\widetilde{u} | \widetilde{\boldsymbol{x}} = \boldsymbol{x}^{[j]}) \cong \mathcal{B}\left(p(\boldsymbol{x}^{[j]}) / c_{\boldsymbol{\lambda}} q_{\boldsymbol{\lambda}}(\boldsymbol{x}^{[j]})\right)$;
3. if $u^{[j]} = 1$, then accept $\boldsymbol{x}^{[j]}$ as a valid sample; otherwise reject it.

This process actually enables sampling from a wide variety of pdfs, differenced by their value of $\boldsymbol{\lambda}$. Since a choice is possible, it is therefore interesting to choose an "optimal" value. The loss function quantifying this interest is given by the probability of acceptance (or, equivalently, one minus the probability of rejection):

Theorem 2 *Consider the rejection scheme previously defined, with $\boldsymbol{\lambda}$ fixed. The probability of acceptance at each step, defined by $\mathrm{p}(\widetilde{u}=1)$, is then given by*

$$\mathrm{p}(\widetilde{u}=1)=\frac{1}{c_{\boldsymbol{\lambda}}}. \tag{5}$$

Proof: Using the marginalization formula, the probability of acceptance $\mathrm{p}(\widetilde{u}=1)$ reads

$$\begin{aligned}\mathrm{p}(\widetilde{u}=1) &= \int \mathrm{p}(\widetilde{u}=1|\widetilde{\boldsymbol{x}}=\boldsymbol{x})\cdot \mathrm{p}(\widetilde{\boldsymbol{x}}=\boldsymbol{x})\,d\boldsymbol{x}\\ &= \int \frac{p(\boldsymbol{x})}{c_{\boldsymbol{\lambda}} q_{\boldsymbol{\lambda}}(\boldsymbol{x})}\cdot q_{\boldsymbol{\lambda}}(\boldsymbol{x})\,d\boldsymbol{x}\\ &= \int \frac{p(\boldsymbol{x})}{c_{\boldsymbol{\lambda}}}\,d\boldsymbol{x}\\ &= \frac{1}{c_{\boldsymbol{\lambda}}},\end{aligned}$$

since $p(\boldsymbol{x})$ is normalized to 1. □

Once the loss function has been clearly defined, its maximization is straightforward:

Theorem 3 *Let $n_0 \in \mathrm{argmax}_{1\leq n\leq N}\left[\sup_{\boldsymbol{x}} f_n(\boldsymbol{x})\right]$. The probability of acceptance $\mathrm{p}(\widetilde{u}=1)$ is then maximized for $\boldsymbol{\lambda}=\boldsymbol{\lambda}^{\max}$ so that $\lambda_{n_0}^{\max}=1$ and $\lambda_{n\neq n_0}^{\max}=0$.*

Proof: From Theorem 2, maximizing the probability of acceptance is equivalent to minimizing $c_{\boldsymbol{\lambda}}$. Since the f_n are pdfs, the α_n are positive, and the following inequalities hold:

$$\lambda_n\alpha_n \geq \lambda_n \min_{1\leq m\leq N}(\alpha_m) \qquad 1\leq n\leq N.$$

Adding up all these inequalities leads to

$$\sum_{n=1}^{N}\lambda_n\alpha_n \geq \sum_{n=1}^{N}\lambda_n \min_{1\leq m\leq N}(\alpha_m)=\min_{1\leq m\leq N}(\alpha_m),$$

since the λ_n sum to 1. $k\min_{1\leq n\leq N}(\alpha_n)$ is therefore a lower bound for $c_{\boldsymbol{\lambda}}$. Now define $n_1 \in \mathrm{argmin}_{1\leq n\leq N}[\alpha_n]$ and $\boldsymbol{\lambda}^{(0)}$ so that $\lambda_{n_1}^{(0)}=1$ and $\lambda_{n\neq n_1}^{(0)}=0$. Then the previous inequalities are in fact equalities. $k\min_{1\leq n\leq N}(\alpha_n)$ is therefore the minimum of $c_{\boldsymbol{\lambda}}$, obtained for $\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(0)}$. But since α_n rereads as

$$\alpha_n=\frac{\prod_{1\leq m\leq N}\sup_{\boldsymbol{x}} f_m(\boldsymbol{x})}{\sup_{\boldsymbol{x}} f_n(\boldsymbol{x})},$$

minimization of α_n tantamounts to maximization of $\sup_{\boldsymbol{x}} f_n(\boldsymbol{x})$. Choice of n_1 is therefore equivalent to the choice of the n_0 of the theorem, and the proof is complete. □

Note that, though highly improbable, the maximum may not be unique, and it is still possible that two or more functions have the same maximum. Any such function will then do.

The issue of sampling from $q_{\boldsymbol{\lambda}}(\boldsymbol{x})$ can now be tackled very easily, since the choice of $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\max}$ greatly simplifies the calculations: $q_{\boldsymbol{\lambda}}(\boldsymbol{x})$ boils down to $f_{n_0}(\boldsymbol{x})$ and $c_{\boldsymbol{\lambda}}$ to $k\alpha_{n_0}$. The modified rejection sampling scheme then rereads

1. sample $\boldsymbol{x}^{[j]}$ according to $\widetilde{\boldsymbol{x}} \cong f_{n_0}(\boldsymbol{x})$;
2. sample $u^{[j]}$ according to $(\widetilde{u}|\widetilde{\boldsymbol{x}} = \boldsymbol{x}^{[j]}) \cong \mathcal{B}(r^{[j]})$ with
$$r^{[j]} \equiv \frac{p(\boldsymbol{x}^{[j]})}{k\alpha_{n_0} f_{n_0}(\boldsymbol{x}^{[j]})} = \prod_{n \neq n_0} \left[\frac{f_n(\boldsymbol{x}^{[j]})}{\sup_{\boldsymbol{x}} f_n(\boldsymbol{x})} \right];$$
3. if $u^{[j]} = 1$, then accept $\boldsymbol{x}^{[j]}$ as a valid sample, otherwise reject it.

Since the normalization constant k does not appear in the procedure, it has the nice consequence that the pdf $p(\boldsymbol{x})$ that we wish to sample from does not need to be normalized.

Approximate acceptance rate. If the pdfs involved in the defining product of $p(\boldsymbol{x})$ are Gaussian with means μ_n and variances σ_n^2 (or, equivalently, concentrations $\upsilon_n^2 = \sigma_n^{-2}$), the resulting pdf $p(\boldsymbol{x})$ is also Gaussian with mean μ and concentration υ^2, such that

$$\mu = \frac{\sum_{n=1}^N \upsilon_n^2 \mu_n}{\sum_{n=1}^N \upsilon_n^2} \qquad \text{and} \qquad \upsilon^2 = \sum_{n=1}^N \upsilon_n^2$$

(see Table 1). In this case, the acceptance rate calculates easily and is equal to

$$\mathrm{p}(\widetilde{u} = 1) = \sqrt{\frac{\upsilon_{n_0}^2}{\sum_{n=1}^N \upsilon_n^2}} \exp\left[\frac{1}{2} \left(\upsilon^2 \mu^2 - \sum_{n=1}^N \upsilon_n^2 \mu_n^2 \right) \right]. \tag{6}$$

Likewise, for multinormal densities with means $\boldsymbol{\mu}_n$ and concentration matrices $\boldsymbol{\Upsilon}_n$, this reads

$$\mathrm{p}(\widetilde{u} = 1) = \sqrt{\frac{|\boldsymbol{\Upsilon}_{n_0}|}{\left|\sum_{n=1}^N \boldsymbol{\Upsilon}_n\right|}} \exp\left[\frac{1}{2} \left(\boldsymbol{\mu}^{\mathrm{t}} \boldsymbol{\Upsilon} \boldsymbol{\mu} - \sum_{n=1}^N \boldsymbol{\mu}_n^{\mathrm{t}} \boldsymbol{\Upsilon}_n \boldsymbol{\mu}_n \right) \right]. \tag{7}$$

In case the pdfs involved are not Gaussian but are unimodal, and if we have access to their means and variances, it is possible to use Equations (6) and (7) as a rule of thumb for the expected acceptance rate. Since a lower acceptance rate requires more drawings to attain a sample of given size, this criterion can in turn be directly related to the computational burden of the procedure for a given configuration.

Note that choice of a function f_n different from the optimal f_{n_0} as comparison function decreases the acceptance rate by a factor $\upsilon_{n_0}/\upsilon_n$ (resp. $\sqrt{|\boldsymbol{\Upsilon}_{n_0}|/|\boldsymbol{\Upsilon}_n|}$). This indicates that the more different the variances are, the more efficient the optimal selection is. This point is further discussed in the simulations.

product	resulting distribution
$\prod_{n=1}^{N} \mathcal{N}(\mu_n, v_n^{-2}; x)$	$\mathcal{N}\left(\frac{\sum_{n=1}^{N} v_n^2 \mu_n}{\sum_{n=1}^{N} v_n^2}, \frac{1}{\sum_{n=1}^{N} v_n^2}; x\right)$
$\prod_{n=1}^{N} \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Upsilon}_n^{-1}; \boldsymbol{x})$	$\mathcal{N}\left(\left(\sum_{n=1}^{N} \boldsymbol{\Upsilon}_n\right)^{-1} \sum_{n=1}^{N} \boldsymbol{\Upsilon}_n \boldsymbol{\mu}_n, \left(\sum_{n=1}^{N} \boldsymbol{\Upsilon}_n\right)^{-1}; \boldsymbol{x}\right)$

Table 1: Stability properties of Gaussian pdfs.

3 Results from simulations

The toy example given in Section 3.1 was developed along three directions to illustrate the influence of various parameters on the sampling efficiency.

Products of Gaussian pdfs were involved in the simulation process. Resulting from the stability properties of these functions (see Table 1), this enabled rigourous control of the sampling procedure, by direct comparison of the statistic summaries resulting from the sampling scheme with the corresponding values of the summaries directly calculated from the true Gaussian pdf.

The programs used for the simulation were written in Matlab and processed on a SunSPARC Ultra 10 workstation.

3.1 Example 1

We chose the product $p(x) \propto f_1(x) \cdot f_2(x)$ with

$$\begin{aligned} f_1(x) &= \mathcal{N}(0, 1; x) \\ f_2(x) &= \mathcal{N}(1, 0.1; x) \end{aligned}$$

In this case, $p(x)$ is known to be Gaussian distributed with mean $\mu = 0.9091$ and variance $\sigma^2 = 0.0909$ from Table 1. The full density function was infered from a 1,000-sample experiment. From Equation (6), 605 samples were expected to be accepted over the 1,000 proposed from the selected comparison function. The run considered took 337 ms to process, and 589 samples were kept. The estimated mean was $\widehat{\mu} = 0.9195$ (1.15% relative error) and estimated variance $\widehat{\sigma^2} = 0.0945$ (3.96% relative error). The estimated pdf is given in Figure 1, showing very good fit with the true pdf.

3.2 Example 2

An important factor is the relative variance of the distributions compared to their relative mean, termed here as "overlapping". In order to test this parameter, we generalized the previous simulation by allowing the variances of both pdfs to vary.

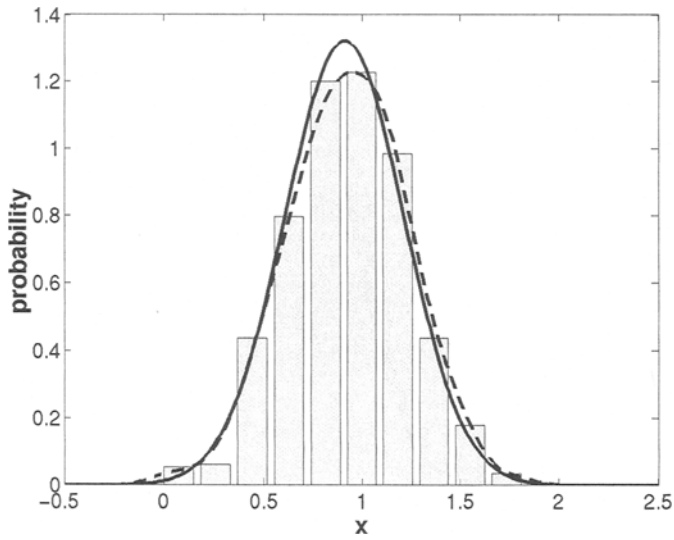

Figure 1: Example 1. Comparison between real pdf (solid line), kernel estimate from a Gaussian kernel of variance 0.01 (dashed line) and histogram.

σ_1^2 / σ_2^2	0.01		0.10		1.00	
0.01	0.0	(0)	10.3	(10)	606.9	(607)
0.10	9.9	(10)	57.9	(58)	606.2	(605)
1.00	607.6	(607)	602.0	(605)	550.3	(551)

Table 2: Example 2. Average number of accepted samples over the 1,000 proposed, among 100 repetitions. The expected numbers are given in parentheses.

More precisely, we chose the product $p(x) \propto f_1(x) \cdot f_2(x)$ with

$$\begin{aligned} f_1(x) &= \mathcal{N}(0, \sigma_1^2, x) \\ f_2(x) &= \mathcal{N}(1, \sigma_2^2, x) \end{aligned}$$

where σ_1^2 and σ_2^2 vary in $\{0.01, 0.1, 1\}$. 100 runs of rejection sampling were performed, each one including 1,000 samples. Average efficiencies and expected numbers of acceptance are compared in Table 2. As hypothesized, overlapping has a dramatic influence on the acceptance rate. This example also clearly shows that the more different the pdf variances are, the more efficient the algorithm is. Finally, the effect of pdf selection appears clearly. When the two pdfs have different variances, choosing the wrong pdf as a comparison function dramatically decreases the acceptance rate, dividing it by a factor of $\sqrt{10}$, resp. $\sqrt{100} = 10$ according to Equation (6). Hence the importance of selecting the optimal function and the corresponding optimal upper bound.

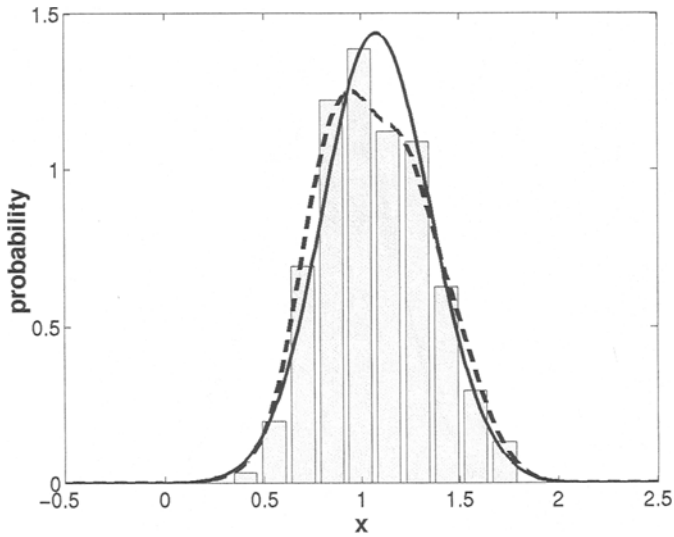

Figure 2: Example 3. Comparison between real pdf (solid line), kernel estimate with a Gaussian kernel of variance 0.01 (dashed line) and histogram.

3.3 Example 3

In order to increase the number of functions in the product, we defined $p(x) \propto f_1(x) \cdot f_2(x) \cdot f_3(x)$ with

$$\begin{aligned} f_1(x) &= \mathcal{N}(0, 1, x) \\ f_2(x) &= \mathcal{N}(1, 0.1, x) \\ f_3(x) &= \mathcal{N}(2, 0.5, x) \end{aligned}$$

Note that the product of the first two functions corresponds to the one studied earlier in Section 3.1, and is a rather favorable case as far as the acceptance rate is concerned, as shown in Section 3.2.

According to Table 1, the resulting pdf is normal distributed with mean $\mu = 1.0769$ and variance $\sigma^2 = 0.0769$. Out of the 1,000 proposals, about 203 are expected to be accepted. This corresponds to one-third of the number of acceptances with only the first two pdfs.

The sampling considered took 525 ms, and 202 samples were kept. From there, the mean was estimated to $\widehat{\mu} = 1.0658$ (-1.03% relative error) and the variance to $\widehat{\sigma^2} = 0.0688$ (-10.53% relative error). Even though the acceptance rate is relatively low, the estimated distribution is fairly good, as shown in Figure 2.

3.4 Example 4

For $\boldsymbol{x} \in \mathbb{R}^2$, we chose the product $p(\boldsymbol{x}) \propto f_1(\boldsymbol{x}) \cdot f_2(\boldsymbol{x})$ with

$$\begin{aligned} f_1(\boldsymbol{x}) &= \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_1\\ \rho_1 & 1\end{pmatrix}, \boldsymbol{x}\right) \\ f_2(\boldsymbol{x}) &= \mathcal{N}\left(\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0.1 & 0.1\,\rho_2\\ 0.1\,\rho_2 & 0.1\end{pmatrix}, \boldsymbol{x}\right), \end{aligned}$$

No correlation ($\rho_1 = \rho_2 = 0$). In this case, the marginals of these densities are exactly the functions defined in Example 1. In this 2-dimensional case, from Table 1, $p(\boldsymbol{x})$ is known to be Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, such that

$$\boldsymbol{\mu} = \begin{pmatrix}0.9091\\0\end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix}0.0909 & 0\\0 & 0.0909\end{pmatrix}.$$

565 samples (compared to an expected number of 577) were kept out of the 1,000 performed from the selected comparison function in 743 ms. The estimated mean $\widehat{\boldsymbol{\mu}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}$ were calculated to

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix}0.9120\\0.0050\end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \begin{pmatrix}0.0848 & 0.0018\\0.0018 & 0.0953\end{pmatrix}.$$

This corresponds to a relative error of 0.64% on the mean and of 7.05% on the variance[1].

Rejection sampling with f_1 as comparison function would decrease the efficiency of the procedure by a factor of 10, hinting that the advantage of optimal selection increases with increasing dimension.

General case. We also allowed ρ_1 and ρ_2 to vary, to show the influence of correlations on the sampling procedure. Whatever the correlation values, f_2 was used as the comparison function. The results are summarized in Table 3. For a given ρ_2, it appars that the acceptance rate increases with decreasing $|\rho_1|$. Another feature is that, when ρ_1 is positive, the algorithm behaves slightly better with a negative ρ_2, and *vice versa*, and the bigger $|\rho_2|$, the bigger the asymmetry.

3.5 Example 5

We finally tested the behavior of ARSP on a problem where non-normal distributions were involved. The following example has also the advantage of showing that, as advocated in the introduction, the form of the distribution assumed in Equation (2) is ubiquitous in Bayesian analysis.

[1] respectively defined as $\frac{\|\widehat{\mu}-\mu\|_2}{\|\mu\|_2}$ and $\frac{\|\widehat{\Sigma}-\Sigma\|_2}{\|\Sigma\|_2}$

ρ_1 / ρ_2	-0.8	-0.4	0	0.4	0.8
-0.8	256.2	284.2	302.7	316.0	327.7
-0.4	530.6	529.1	532.4	531.4	531.7
0	575.7	576.5	577.8	577.1	577.6
0.4	534.0	530.5	534.2	531.5	526.1
0.8	325.0	316.6	303.2	284.6	255.3

Table 3: Example 3. Average number of accepted samples over the 1,000 proposed, among 100 repetitions.

For this, we simulated a vector $\boldsymbol{y}$ of $N = 10$ i.i.d. samples generated by a normal distribution with 0 mean and unknown variance σ^2 that had to be inferred in a Bayesian framework. We further assumed that we had about σ^2 a vague prior, namely that σ^2 was around 1 with a rather large uncertainty. This information was in turn translated into a Gamma pdf:

$$p(\sigma^2) = \Gamma(4, 4; \sigma^2).$$

Relating the posterior pdf with the likelihood and the prior pdf through Bayes' theorem then led to

$$\begin{aligned} p(\sigma^2|\boldsymbol{y}) &\propto p(\sigma^2) \cdot p(\boldsymbol{y}|\sigma^2) \\ &= \Gamma(4, 4; \sigma^2) \cdot \chi_8^{-2}\left(\frac{\boldsymbol{y}^t\boldsymbol{y}}{8}; \sigma^2\right). \end{aligned}$$

The mode of the Gamma pdf was 0.896 (reached for $\sigma^2 = 0.75$), which was smaller than the maximum of the scaled inverse-chi-square, which was 1.204 (for $\sigma^2 = 0.580$) with our dataset. Rejection sampling was hence performed with the scale inverse-χ^2 pdf as comparison function. Out of the 1,000 samples drawn from this pdf, Equation (6) predicted that 808 samples would be kept, and 739 actually were. The results, shown in Figure 3, exhibit very good fit of the approximation.

4 Discussion

With increasing use of Markov Chain Monte Carlo methods, faster, simpler and more efficient methods for generating exact multidimensional random samples are required. In this paper we have proposed and illustrated a new method, called Automated Rejection Sampling from Product of distributions. ARSP is applicable to all problems for which the distribution has the form of a product of standard pdfs. Products of distributions are ubiquitous in Bayesian analysis, and ARSP has therefore a wide range of application.

The striking advantage of our method is that it avoids the biggest difficulty associated with the conventional black-box algorithm by automatically identifying the majorating pdf and calculating the optimal acceptance rate, given by maximization

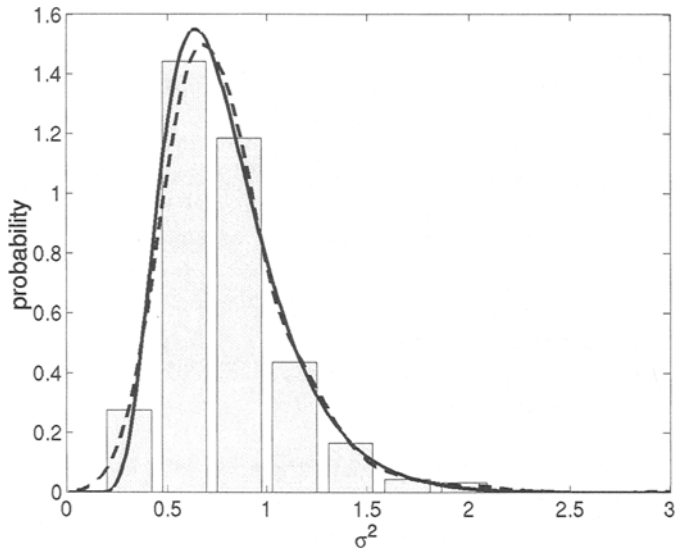

Figure 3: Example 5. Comparison between real pdf (solid line), kernel estimate with a Gaussian kernel of variance 0.01 (dashed line) and histogram.

of the probability of acceptance. The only practical limitation of this method resides in the computational power one has at his or her disposal, and consequently the time required to draw a sample of given size. The examples developed suggest that this efficiency greatly varies with regard to the number of functions involved in the product as well as the intrinsic properties of these functions. However, case-by-case estimation of the computational burden can be approximatively performed through the rule of thumb given by Equations (6) and (7). Finally, ARSP seems to be suited to design universal algorithms for a very large class of pdfs. In this context, some requirements are needed to improve the method.

Product reduction. To derive a universal algorithm from ARSP, the very first step consists of simplifying the product of distributions, so as to keep as few terms as possible in the product (*e.g.* using Table 1). As a consequence, this will mecanically increase the acceptance rate.

Efficiency and pdfs. The form of the pdfs involved has a deep impact on the efficiency of the procedure. For instance, Cauchy distributions revealed practically intractable in our simulations for a product of more than five such functions, since the number of samples to draw from the comparison density to get 100 samples accepted was greater than 10^7 and overflew the memory. This is a consequence of the fact that Cauchy distributions have wide tails. A sample originating from one Cauchy pdf is therefore very likely to have a low probability for the other pdfs,

leading to a vanishing acceptance ratio.

Efficiency and information coherence. Efficiency is usually linked with the quality of the approximation pdf. In our case, this directly relates to the coherence of the information brought by the different pdfs of the product relative to each other. This idea of shared information can be fully understood in the light of simulation example 5, where one pdf is a likelihood function, and the other a prior pdf. The more the prior and the likelihood locate the parameter precisely and accordingly, the better the procedure will perform. On the other hand, if both pdfs involved in the product defining $p(\boldsymbol{x})$ separately estimate the variable $\boldsymbol{x}$ very precisely but disagreeingly (*i.e.* if they are peaked around very different values), the acceptance rate will be very low. In this respect, the case presented in example 1 with $\sigma_1^2 = \sigma_2^2 = 0.01$ is pathologically extreme, and we believe that this situation does not occur often in real-life problems.

In one dimension, when the expected acceptance ratio is really low, it might be more convenient to resort to other methods, such as ARS. However, all other methods require extra information that we do not assume to be available here and that cannot easily be inferred from the information at hand, such as the mode of the product. It is therefore not obvious that the time needed by a competitive algorithm to extract this information will counterbalance the time required for a few thousand more tries in ARSP.

In higher dimension, ARSP is, to our knowledge, the only method that allows for automated multi-dimensional sampling, and its performances are again closely related to the informational coherence between densities of the product. In case it is applied to achieve a step of Gibbs sampling, a disagreement between pdfs is most likely to appear at the beginning, during the burn-out period, when the current state is still highly dependent of the randomly chosen seed. Choice of another seed would then eliminate the problem and allow for efficient sampling.

We finally advocate that this weakness of the procedure can also be considered as an interesting feature in Bayesian model evaluation. Indeed, a very low acceptance rate is typical of a strong disagreement between information originating from different parts of the model, somehow revealing a hidden inconsistency. Models leading to very poor acceptance rates should be examinated with great care, since the pathological behavior of ARSP could as well be the computational translation of a pathological or contradictory model.

5 Conclusion

We have proposed an efficient rejection method for automated generation of exact multidimensional random samples. The striking feature of our method is that it avoids some difficulties associated with conventional algorithms by automatically identifying the majoring pdf and calculating the optimal acceptance rate given by maximization of the probability of acceptance. The only requirement is that the sampled pdf be proportional to a product of standard distributions. The utility of

the ARSP method was illustrated with various simulations, regarding the number of distributions in the product and the dimensionality of the variate. Further research includes assessment of its practical benefits, and especially compared to other automated sampling procedures.

6 Acknowledgments

The authors are grateful to M. Pélégrini-Issac for her kind proofreading the paper. G. Marrelec is supported by the Fondation pour la Recherche Médicale.

References

Dagpunar, J. (1988). *Principles of Random Variate Generation*. Oxford: Clarendon Press.

Devroye, L. (1986). *Nonuniform Random Variate Generation* (John Wiley and Sons ed.). Springer, New York.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1998). *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall, London.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampler for Gibbs sampling. *Journal of Applied Statistics 41*, 337–348.

Kinderman, A. J. and F. J. Monahan (1977). Computer generation of random variable using the ratio of uniform deviates. *ACM Transactions on Mathematical Software 3*(3), 257–260.

Leydold, J. (1998). A rejection technique for sampler from log-concave multivariate distributions. *ACM Transactions on Modeling and Computer Simulation 8*(3), 254–280.

Neal, R. M. (2000). Slice sampling. Technical report, Department of Statistics, University of Toronto.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C: The Art of Science Computing*. Cambridge University Press.

Ruanaidh, J. J. K. O. and W. J. Fitzgerald (1996). *Numerical Bayesian Methods Applied to Signal Processing*. Statistics and Computing. Springer, New York.

Tanner, M. A. (1994). *Tools for Statistical Inference – Methods for the Exploration of Posterior Distributions and Likelihood Functions* (2nd ed.). Springer Series in Statistics. Springer, New York.

Winkler, G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods – A Mathematical Introduction*. Number 27 in Applications of Mathematics. Springer, Berlin.