Tinbergen Institute Discussion Paper TI 98-071/4

Adaptive Polar Sampling: A New MC Technique for the Analysis of Ill-behaved Surfaces

LUC BAUWENS, CHARLES S. BOS* and HERMAN K. VAN DIJK

CORE, Université catholique de Louvain and

Econometric and Tinbergen Institutes, Erasmus University Rotterdam

August 25, 1998, rev. B

Abstract

Adaptive Polar Sampling is proposed as an algorithm where random drawings are directly generated from the target function (posterior) in all-but-one directions of the parameter space. The method is based on the mixed integration technique of van Dijk, Kloek & Boender (1985) but extends this one by replacing the one-dimensional quadrature step by Monte Carlo simulation from this one-dimensional distribution function. The method is particularly suited for the analysis of ill-behaved surfaces. An illustrative example shows the feasibility of the algorithm.

Keywords: Markov Chain Monte Carlo sampling; Ill-behaved surfaces; Polar coordinates

[*] Correspondence to Charles S. Bos, Tinbergen Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands. Email: cbos@few.eur.nl. Support from HCM grant ERBCHRXCT 940514 is gratefully acknowledged.

1 Introduction

Sampling from a general posterior distribution with density function $p(\theta)$ of a parameter vector θ has been the topic of intensive research in recent years. For well-behaved cases, with the full conditionals given or with the probability mass of the density lying in one, joined region, where the distribution is regular enough, several efficient algorithms are available. The most important sampling methods are the Metropolis-Hastings algorithm, importance sampling and the Gibbs sampler (see e.g. Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953), Kloek & van Dijk (1978) and Smith & Gelfand (1992)).

For more difficult cases, sometimes an algorithm as the Griddy Gibbs sampler (Ritter & Tanner 1992), the local adaptive importance sampler of Givens & Raftery (1996) or the simulated tempering or sintering algorithms (see Liu & Sabatti (1998)) may lead to a solution. In this paper we propose adaptive polar sampling (APS) as an alternative. This algorithm can be used in less well-behaved cases, which may exhibit multiple modes, ridges in the density function, non-Gaussian tail behavior or e.g. strong correlation between the elements of the parameter vector.

2 Basics of adaptive polar sampling

Let the joint density function $p(\theta)$ of the parameter vector θ be of a known parametric form. Assume that the probability mass lies in a bounded region $R \in \Re^n$. In case the bounds are not known, take them 'large' such that every area of interest is contained within R. Suppose μ and Σ are preliminary estimates of the location and scale. As a first step, standardize the parameter vector θ using the location and scale as

$$y = \Sigma^{-1/2}(\theta \Leftrightarrow \mu)$$

with $\Sigma^{1/2}$ the Cholesky decomposition of the scale matrix Σ.

The probability mass in terms of the parameterization y can be expected to be distributed more or less evenly in all directions around the origin of the space. The distribution of the distance of the mass from the origin may still behave quite erratically.

Given this observation, we propose to transform the n-dimensional vector y into polar coordinates (see Kendall & Stuart (1969), chapter 11 for details), leaving an $n \Leftrightarrow 1$ dimensional vector η of directions and a scalar ρ indicating the length of the vector y. As will be shown in the next section, the marginal distribution of η is expected to be reasonably well behaved. In order to sample from this marginal distribution using Metropolis-Hastings, the marginal density of η for the case where θ is assumed to be normally distributed can be used as the candidate density. Given a direction η, a corresponding element ρ can be sampled from the univariate conditional distribution function, by inverting the cumulative distribution function of ρ in the direction η. In this decomposition of the parameter space into polar coordinates the similarity with the Mixed Integration algorithm of van Dijk et al. (1985) is found.

After collecting a sample conditional on μ and Σ, these estimates can be adapted using the first and second moment of the training sample. In a new rotation, the updated estimates can be used in order to get a better behaved marginal distribution of η.

3 Transformed densities

Denote the transformation of θ into (η, ρ) by $(\eta, \rho) = T(\theta)$. We suppress the dependence of the transformation on μ and Σ for notational convenience. The density in terms of the polar coordinates can be written as

$$p_{\eta,\rho}(\eta,\rho) = p_\theta \Big/ T^{-1}(\eta,\rho)\big) \, |J(\eta,\rho,\Sigma)|$$

with a Jacobian

$$J(\eta, \rho, \Sigma) = \det(\Sigma)^{-\frac{1}{2}} \rho^{n-1} \left(\prod_{i=1}^{n-2} \cos^{n-i-1} \eta_i \right).$$

The APS method transforms the joint density of η and ρ into

$$p_{\eta,\rho}(\eta, \rho) = p_{\rho \mid \eta}(\rho \mid \eta) p_\eta(\eta)$$

where

$$p_\eta(\eta) = \int_{r(\eta \mid R)} p_\theta \ \not{T}^{-1}(\eta, \rho)) \left| J(\eta, \rho, \Sigma) \right| \partial\rho.$$

Here $r(\eta \mid R)$ is the region in terms of ρ in the direction η as prescribed by the region of interest R in the original metric.

It is illustrative to apply this transformation to polar coordinates to a variable which is distributed as a bivariate normal, $\theta \sim \mathcal{N}(\mu, \Sigma)$. Then we get $y \sim \mathcal{N}(\mathbf{0}, I_2)$, $\rho = \sqrt{y_1^2 + y_2^2}$ $(\rho \geq 0)$ and η is the unique solution of $\cos \eta = y_1/\rho$ and $\sin \eta = y_2/\rho$ $(\eta \in [0, 2\pi))$. The joint, marginal and conditional distributions of interest are

$$\begin{aligned} p_{\eta,\rho}(\eta, \rho) &\propto \rho \; e^{-\rho^2/2} \\ p_\eta(\eta) &= \frac{1}{2\pi} \\ p_{\rho \mid \eta}(\rho \mid \eta) &\propto \rho \; e^{-\rho^2/2} \quad \Leftrightarrow \quad \rho^2 \mid \eta \sim \exp(1/2). \end{aligned}$$

For the bivariate normal distribution function the marginal density of the (univariate) directions η is uniform. In the general case, for the multivariate normal distribution, the last direction η_{n-1} is uniformly distributed between 0 and 2π, whereas the other directions $\eta_1, .., \eta_{n-2}$ have a marginal density $p(\eta_i) \propto \cos^{n-i-1} \eta_i$ $(\eta_i \in [\Leftrightarrow\pi, \pi),\ i = 1, .., n \Leftrightarrow 2)$.

For other, non-normal distribution functions the marginal of η at most approximately follows this combined uniform/cosine density. The difference, which depends on the case studied, is not expected to be extremely large. Sampling from the marginal distribution of η using a Metropolis-Hastings step with the uniform/cosine density as a candidate function is expected to give reasonable acceptance rates.

4 The APS algorithm: Sampling by parts

The similarity of the marginal distribution of the direction η to the marginal distribution in the case the distribution of θ is normal is exploited in a Metropolis-Hastings sampling step. Assume we have a starting value $\eta^{(0)}$, and have put $i = 1$. Then the algorithm proceeds in drawing $\eta^{(i)}$'s as follows:

- Draw $\eta^* = (\eta_1^* \dots \eta_{n-1}^*)$ from the candidate density corresponding to the normal density for θ as it was indicated above. Denote this candidate density by $\tilde{p}_\eta$. The sampling is most easily done by drawing $y^* \sim \mathcal{N}(0, I_n)$ and transforming this y^* into polar coordinates (ρ^*, η^*), where only η^* is used in the next steps.
- Calculate
$$\alpha = \min \left\{ \frac{p_\eta(\eta^*)}{\tilde{p}_\eta(\eta^*)} \Bigg/ \frac{p_\eta(\eta^{(i-1)})}{\tilde{p}_\eta(\eta^{(i-1)})}, 1 \right\}$$
with $\eta^{(i-1)}$ the η drawn in the previous iteration,
- Take $\eta^{(i)} = \begin{cases} \eta^* & \text{with probability } \alpha \\ \eta^{(i-1)} & \text{with probability } 1 \Leftrightarrow \alpha. \end{cases}$
- If desired, increase i and repeat the above steps.

In this sampler for η the density function in the direction of η is explored while calculating the marginal $p_\eta(\eta)$. With this information, it is no large effort to sample $\rho^{(i)} \,|\, \eta^{(i)}$ from the inverse distribution function, as is done e.g. in the Griddy Gibbs sampler (see Ritter & Tanner

(1992)). The sampled values $(\eta^{(i)}, \rho^{(i)})$ are then transformed back into the original metric using the inverse transformation, $\theta^{(i)} = T^{-1}(\eta^{(i)}, \rho^{(i)})$.

The sampling is done under the assumption that the location and scale parameters μ and Σ, used in the transformation $\theta \to (\eta, \rho)$, are known. When this is not the case (that is, usually), the algorithm can be used with a rough, preliminary estimate of these parameters; one or two small rounds of sampled values can be used to improve on the estimate of μ and Σ, before the algorithm is run for collecting the final sample. We note that poor estimates of location and scale do not lead to an incorrect distribution of the sampled values, only the acceptance rate in the sampling of η's is expected to be lower.

5 An illustration on an ill-behaved density function

The focus of several papers in the MCMC literature is directed at the sampling from mixture distributions. For these distributions Gibbs sampling may not converge if the probability of switching from one mode to the other is close to zero. When using Metropolis-Hastings, it can be hard to find a good, general candidate density function which is able to cover all sorts of mixtures. Very low acceptance rates and strong correlation in the sampling output are often the results.

The model we put to the test is the following. Take

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \begin{cases} \mathcal{N}(\mu_1, \Sigma_1) & \text{with probability } p \\ \mathcal{N}(\mu_2, \Sigma_2) & \text{with probability } 1 \Leftrightarrow p \end{cases}$$

This is a canonical example of a bivariate bimodal mixture model. As long as the modes are not connected, and they do not lie next to each other in the direction of one of the axes, Gibbs sampling does not converge on this problem. A general candidate function for Metropolis-

Table 1: Results of location and scale estimates in successive samples

Sample	$\#\eta$	$\frac{\#\rho}{\#\eta}$	Acc. rate η	Location $\mu^{(\cdot)}$	Scale $\Sigma^{(\cdot)}$
0				$\begin{pmatrix} 5 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$
1	100	50	1	$\begin{pmatrix} 3.4 \\ 3.1 \end{pmatrix}$	$\begin{pmatrix} 15.6 & \Leftrightarrow 2.6 \\ \Leftrightarrow 2.6 & 16.7 \end{pmatrix}$
2	100	50	.18	$\begin{pmatrix} \Leftrightarrow 0.5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} 17.1 & \Leftrightarrow 15.9 \\ \Leftrightarrow 15.9 & 17.9 \end{pmatrix}$
3	10000	5	.16	$\begin{pmatrix} 0.0 \\ 0.0 \end{pmatrix}$	$\begin{pmatrix} 17.5 & \Leftrightarrow 16.1 \\ \Leftrightarrow 16.1 & 17.6 \end{pmatrix}$

Hastings is hard to find, especially if the two modes are not close.

In our illustration we choose $\mu_1 = (4, \Leftrightarrow 4)'$, $\Sigma_1 = I_2$, $\mu_2 = (\Leftrightarrow 4, 4)'$, $\Sigma_2 = 2I_2$ and $p = 1/2$, with I_2 the 2×2 identity matrix. The preliminary 'estimates' of the location and the scale of $\theta = (x, y)'$ are taken as $\mu^{(0)} = (5, 5)'$ and $\Sigma^{(0)} = I_2$. These initial values are far away from the true location and scale of the density of $\mu^{(\text{true})} = (0, 0)'$ and $\Sigma^{(\text{true})} = \begin{pmatrix} 17.5 & -16.0 \\ -16.0 & 17.5 \end{pmatrix}$. The region of interest was chosen as $R = [\Leftrightarrow 15, 15] \times [\Leftrightarrow 15, 15]$.

To improve on the preliminary estimates of location and scale, two rotations of the algorithm were used. We sampled 50 values ρ in 100 different directions η. In constructing the first training sample we skipped the Metropolis-Hastings step on η, to get a very quick first improvement on $\mu^{(0)}$ and $\Sigma^{(0)}$. The second rotation used MH to sample from the true distribution of η, resulting in an acceptance rate of η of 0.18. From this training sample the location and scale parameters $(\mu^{(2)}, \Sigma^{(2)})$ were estimated, to be used in drawing the final sample of 10000 different values of η. The final acceptance rate of η was 0.16, as is reported in Table 5. In this final rotation, about 16% of the sampled values of η where accepted.

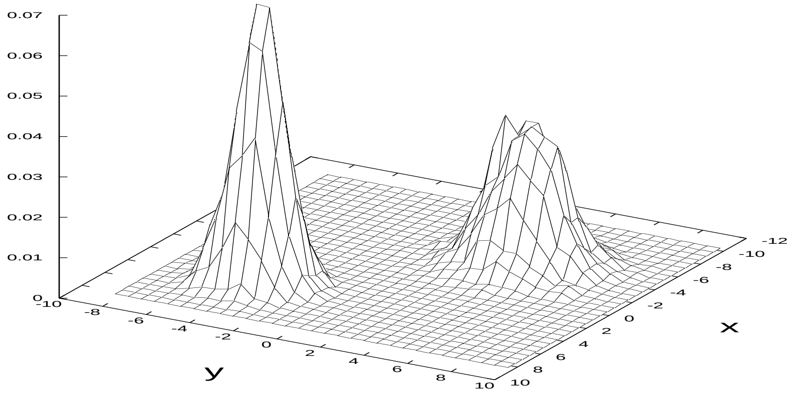

Figure 1: Results for the joint density of the two elements of the bivariate normal mixture distribution

The algorithm continued until 10000 η's were accepted, resulting in a total sample of size $10000 \times 5/0.16 \approx 310000$ elements of θ. Figure 1 depicts the joint density of the sampled θ's, where a kernel smoother was applied in constructing the graph. In Figure 2 the marginal distributions of the sampled x and y are shown.

Implementation of the algorithm was done in the Gauss programming language. The run which was described above took about 15 minutes on a Pentium 233Mhz.

6 Conclusions

An alternative method for sampling from a posterior density has been introduced. This Adaptive Polar Sampling is a new MC algorithm which is better able to handle ill-behaved surfaces,

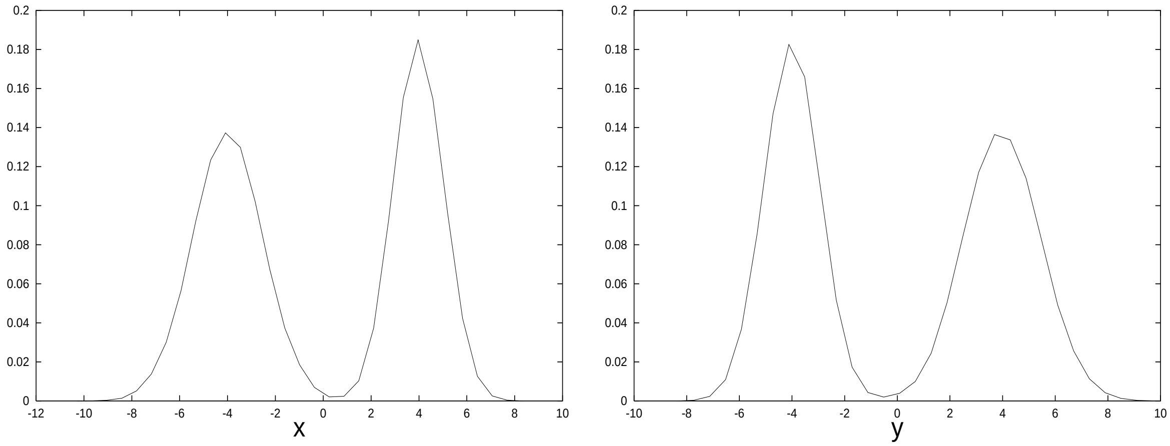

Figure 2: Results for the marginal densities of the two elements of the bivariate normal mixture distribution

where other sampling algorithms may exhibit slow convergence or may not converge at all. This greater flexibility comes at the cost of having to calculate a one-dimensional integration at every step during the sampling. For irregular posterior densities however the greater flexibility and better expected convergence rate will outweigh the cost of the integration.

As an illustration we used the algorithm to sample from the bivariate mixture with two modes lying far apart (see also Liu & Sabatti (1998)) who applied the sintering algorithm for sampling from the same model). A sample from the true posterior density was found even though initial estimates of the location and scale parameters used in the transformation were wildly wrong. Adaptation of these parameters in two preliminary steps led to good results. A set of other densities was put to the test as well, but for reasons of space those results are not extensively reported here. In short, results for a tobit, bimodal tobit, ARMA(1,1) with near root cancellation, ARMA(1, 1)-ARCH(1) and for a bivariate uniform density were similar: Two or three training samples where sufficient to get reasonably high acceptance rates in the final sample from the true posterior. Acceptance rates of η usually varied between 30% and

80%. The exception was the bimodal mixture distribution presented here: A larger distance between the modes results in a lower acceptance rates, as the covariance structure of the (training) sample does not describe well the distribution of the density mass in the directions η.

Future work on the improvement of the algorithm can be directed at several topics. First, the problem of convergence should be targeted further. What convergence measures are available for this specific algorithm? The algorithm does not depend strongly on the estimates of location μ and scale Σ that are used in the transformation. Even so, the optimal number and type of preliminary rotations for improving the estimates of these parameters can be investigated more deeply.

The Mixed Integration algorithm (van Dijk et al. 1985), which is based on a similar transformation as the one applied here, uses antithetic sampling as a variance reduction technique. A similar extension of the present algorithm may have further positive effects on the robustness, especially in case the location and scale parameters in the transformation are not accurate, and on the degree of correlation between successive drawings.

At present, the parameters μ and Σ of the algorithm are estimated using the first and second moment of the sample generated by the algorithm in a (preliminary or final) rotation. As the effort of calculating $p_{\rho\,|\,\eta}(\rho\,|\,\eta)$ is made already, it is more efficient if this information was used in a type of Rao-Blackwell estimate of these parameters.

The accuracy of the one-dimensional numerical integration in the calculation of $p_{\eta}(\eta)$ is another topic of future investigation. The way the size of the region of interest R or the presence of strong singularities in the posterior density affects the integration and convergence results needs some practical and theoretical attention as well.

Finally, the algorithm should be tested on models of empirical interest.

References

Givens, G. H. & Raftery, A. E. (1996), 'Local adaptive importance sampling for multivariate densities with strong nonlinear relationships', *Journal of the American Statistical Association* **91**, 132–141.

Kendall, M. G. & Stuart, A. (1969), *The Advanced Theory of Statistics*, Vol. 1: Distribution Theory, London: Griffin.

Kloek, T. & van Dijk, H. K. (1978), 'Bayesian estimates of equation system parameters: An application of integration by monte carlo', *Econometrica* **46**, 1–20.

Liu, J. S. & Sabatti, C. (1998), 'Simulated sintering: Markov Chain Monte Carlo with spaces of varying dimensions', *Proceedings of the 6th International Meeting on Bayesian Statistics, Valencia, Spain* .

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equations of state calculations by fast computing machines', *Journal of Chemical Physics* **21**, 1087–1091.

Ritter, C. & Tanner, M. A. (1992), 'Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler', *Journal of the American Statistical Association* **87**, 861–868.

Smith, A. F. M. & Gelfand, A. E. (1992), 'Bayesian statistics without tears: A sampling-resampling perspective', *The American Statistician* **46**, 84–88.

van Dijk, H. K., Kloek, T. & Boender, C. G. E. (1985), 'Posterior moments computed by mixed integration', *Journal of Econometrics* **29**, 3–18.