Bayesian Analysis (2013) **8**, Number 1, pp. 23–26

Comment on Article by Schmidl et al.

Dawn B. Woodard *

The authors develop a novel Markov chain method (ACIMH) that is designed to sample efficiently by adapting to the features of the target distribution, learning from the samples obtained previously. Their approach is based on independence Metropolis-Hastings (IMH), and takes the proposal distribution to be an estimate of the target distribution. Its appeal is that the efficiency of IMH is controlled by how close the proposal density is to the target density. If a very accurate estimate of the target density can be obtained, then the samples obtained by the Markov chain are nearly independent, leading to near-optimal accuracy of Monte Carlo approximations.

The authors' estimate of the target density is based on D-vine copulas, an extremely flexible class of models that has not been used previously for this purpose. Other authors have proposed similar "adaptive" IMH methods based on alternative estimates of the target density, in particular mixture distributions such as normal or t mixtures (Andrieu and Moulines 2006; Andrieu and Thoms 2008). Continued development of efficient general-purpose sampling algorithms like these is critical as we create robust software packages for Bayesian statistics that will encourage its widespread use.

D-vine copulas use a factorization approach that may scale more effectively with dimension than the mixture model approach. ACIMH factorizes the target density as

$$f(x) = [\prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,(i+j)|(i+1):(i+j-1)}] \cdot [\prod_{k=1}^{d} f_k(x_k)] \qquad x \in \mathcal{X} \tag{1}$$

then drops the dependence of $c_{i,(i+j)|(i+1):(i+j-1)}$ on $x_{(i+1):(i+j-1)}$, so that only univariate and bivariate densities need to be estimated. However, I show that as ACIMH is currently defined its efficiency still degrades exponentially in the dimension d. I do this for several simple but representative target densities. So ACIMH is expected to be ineffective for high-dimensional problems; this may explain why the examples used by Schmidl et al. are low-dimensional (having $d = 2$, 3, and 7). I will argue that this issue is inherent to any IMH method that takes the proposal density to be an estimate of the joint target density based on past samples. I then suggest that this problem may be mitigated by applying ACIMH or a related method to blocks of parameters separately rather than to the entire parameter vector simultaneously. ACIMH is a promising addition to the Markov chain toolbox, due to its ability to flexibly estimate aspects of the target density; however, it needs to be used in a blocked fashion in order to scale well with dimension.

*School of Operations Research & Information Engineering and Department of Statistical Science, Cornell University, Ithaca, NY. http://people.orie.cornell.edu/woodard

© 2013 International Society for Bayesian Analysis DOI:10.1214/13-BA801A

1 Efficiency of ACIMH

Here I analyze the efficiency of ACIMH as a function of the dimension d. ACIMH is not an adaptive MCMC method in the purest sense, namely that the chain continues to adapt forever. ACIMH stops updating the copula after a fixed number of iterations, so one can view it instead as a standard Markov chain with a tuning period. Viewing it in this way, I will give results on the convergence rate of the Markov chain after this period. Take the number of iterations n in the tuning period to be fixed (not dependent on d), and let $\hat{f}(\cdot)$ be the estimator of $f(\cdot)$ obtained at the end of this period.

I argue that: (a) the accuracy of $\hat{f}$ degrades exponentially in d, in the sense that $\inf_{x\in\mathcal{X}} \frac{\hat{f}(x)}{f(x)}$ decays exponentially; and (b) this implies that the convergence rate of IMH with proposal density $\hat{f}$ decays exponentially in d. I will show that (a) holds for several simple but non-pathological target densities. The step (b) is proven by Liu (1996), although the continuous-state-space case is handled incompletely. Specifically, Liu (1996) shows that for a discrete state space the spectral gap (convergence rate) of IMH is equal to $\inf_{x\in\mathcal{X}} \frac{q(x)}{f(x)}$ where $q(\cdot)$ is the proposal distribution. In the continuous-state-space case, which is more technically challenging, he gives evidence strongly suggesting that this result still holds. Combined with (a), this suggests that the spectral gap of ACIMH decays exponentially in d. Informally, this means that the number of iterations of ACIMH required to attain a fixed accuracy increases exponentially in d. More formally, the number of iterations required to decrease the (χ^2) distance to the stationary distribution by a fixed factor grows exponentially in d, in the worst case over starting distributions (cf. Woodard, Schmidler, and Huber 2009). Similar implications hold regarding the accuracy of Monte Carlo estimators.

To show (a) for several examples, I rely solely on the error introduced by estimation of the term $\prod_{k=1}^{d} f_k(x_k)$ in (1). I assume that the bivariate densities $c_{i,(i+j)|(i+1):(i+j-1)}$ satisfy the modeling assumption (do not depend on $x_{(i+1):(i+j-1)}$) and are estimated with perfect accuracy; taking into account this source of error would only increase the overall error. Specifically, I will take target densities that have the product form $f(x) \triangleq \prod_{k=1}^{d} f_k(x_k)$ for $x \in \mathcal{X}$, and assume that the bivariate densities $c_{i,(i+j)|(i+1):(i+j-1)}$ are correctly estimated to be equal to one on the support, so that the goal is to estimate $f(x) = \prod_{k=1}^{d} f_k(x_k)$ by $\hat{f}(x) = \prod_{k=1}^{d} \hat{f}_k(x_k)$. In this simplified context ACIMH can be defined on any state space, not just the Euclidean spaces which are needed for the full copula representation. All that is needed is a parametric form for $\hat{f}_k(x_k)$, the parameters of which are estimated by maximum likelihood as recommended by Schmidl et al. For simplicity I will also assume that the samples $x_i = (x_{i1}, \ldots, x_{id})$ for $i \in \{1, \ldots, n\}$ from the tuning (adaptation) period are i.i.d. according to $f(\cdot)$; taking into account their autocorrelation would only inflate the error.

First, consider the discrete state space $\mathcal{X} = \{0,1\}^d$ and the target density $f(x) \triangleq \prod_{k=1}^{d} f_k(x_k)$ where $f_k(x_k) \triangleq p_k^{x_k}(1-p_k)^{1-x_k}$ and $p_k \in (0,1)$. The maximum likelihood estimator of p_k is $\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$, for $k \in \{1, \ldots, d\}$. To avoid a degenerate estimate replace with $\frac{1}{n}$ if $\sum_i x_{ik} = 0$ and $\frac{n-1}{n}$ if $\sum_i x_{ik} = n$.

Say that the true values of p_k are all $p_k = \frac{1}{2}$, and define

$$W_d \triangleq \ln \frac{\hat{f}(1,\ldots,1)}{f(1,\ldots,1)} = \sum_{k=1}^{d} [\ln \hat{f}_k(1) - \ln f_k(1)] = \sum_{k=1}^{d} [\ln \hat{p}_k - \ln \frac{1}{2}].$$

The quantity $E(\ln \hat{p}_k)$ does not depend on k or d, and by Jensen's inequality $E(\ln \hat{p}_k) < \ln E\hat{p}_k = \ln \frac{1}{2}$, so $EW_d = cd$ for some $c \in (-\infty, 0)$. Also, $\mathrm{Var}(W_d) = d\,\mathrm{Var}(\ln \hat{p}_1)$ where $\mathrm{Var}(\ln \hat{p}_1)$ does not depend on d. By Chebyshev's inequality,

$$\Pr(W_d \geqslant cd^{1/4}) \leqslant \Pr(|W_d - EW_d| \geqslant -cd^{3/4}) \leqslant \frac{d\,\mathrm{Var}(\ln \hat{p}_1)}{c^2 d^{3/2}} \overset{d\to\infty}{\longrightarrow} 0.$$

So $\Pr(\hat{f}(1,\ldots,1)/f(1,\ldots,1) < \exp\{cd^{1/4}\}) \overset{d\to\infty}{\longrightarrow} 1$, meaning that $\inf_{x\in\mathcal{X}} \hat{f}(x)/f(x)$ decays exponentially in d.

For a continuous-state-space example, take $\mathcal{X} = \mathbb{R}^d$, $f(x) = \prod_{k=1}^{d} f_k(x_k)$ and $f_k(x_k) = N(x_k; \mu_k, 1)$. The maximum likelihood estimator of μ_k is $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$. Say that the true values are $\mu_k = 0$, and define $W_d \triangleq \ln \frac{\hat{f}(0,\ldots,0)}{f(0,\ldots,0)} = \sum_{k=1}^{d} [-\frac{1}{2}\hat{\mu}_k^2]$. We have $E\hat{\mu}_k^2 > (E\hat{\mu}_k)^2 = 0$, and the rest of the argument is analogous to the first example.

Here I relied only on the error associated with estimating $f_k(x_k)$, and the fact that it combines multiplicatively when estimating $\prod_{k=1}^{d} f_k(x_k)$. This is not specific to ACIMH; attempting to estimate the joint density $f(x)$ directly without any factorization assumptions would presumably lead to even higher error. So the inefficiency we have identified is inherent to any IMH method that takes the proposal density to be an estimate $\hat{f}(\cdot)$ of the joint density based on previous samples.

2 Conclusion

Although the authors focus on updating the entire parameter vector at once, when d is large it may be more efficient to apply ACIMH to blocks of parameters. The blocks would be chosen to contain highly dependent sets of parameters, and could overlap. Specifically, one would select subsets $A_j \subset \{1,\ldots,d\}$ of the parameter index set, for $j \in \{1,\ldots,J\}$ where J is the desired number of blocks and $\cup_{j=1}^{J} A_j = \{1,\ldots,d\}$. After an initial sampling period, the samples would be used to estimate the marginal density $f_j(x_{A_j})$ of each subvector of the parameters, for instance with vine-copulas. Then one would simulate a Metropolis-within-Gibbs chain, updating the subvectors x_{A_j} in turn according to IMH moves with proposal density $q_j(x_{A_j}) = \hat{f}_j(x_{A_j})$ and acceptance rate $\min\left\{1, \frac{f(x^{\mathrm{new}})q_j(x_{A_j}^{\mathrm{old}})}{f(x^{\mathrm{old}})q_j(x_{A_j}^{\mathrm{new}})}\right\}$. This chain would be more efficient if each proposal density q_j were equal to the *conditional* density $f(x_{A_j}|x_{\{1,\ldots,d\}\setminus A_j})$ of x_{A_j} given the remainder of the parameter vector, since then the acceptance rate would always be equal to one. However, estimating the conditional density $f(x_{A_j}|x_{\{1,\ldots,d\}\setminus A_j})$ would suffer from the same difficulties as estimating the joint density $f(x)$, so I am instead suggesting the use of the estimated *marginal* density $\hat{f}_j(x_{A_j})$. Although suboptimal, this substitution may still yield an efficient algorithm.

References

Andrieu, C. and Moulines, E. (2006). "On the ergodicity properties of some adaptive MCMC algorithms." *Annals of Applied Probability*, 16: 1462–1505. 23

Andrieu, C. and Thoms, J. (2008). "A tutorial on adaptive MCMC." *Statistics and Computing*, 18: 343–373. 23

Liu, J. S. (1996). "Metropolized independent sampling with comparisons to rejection sampling and importance sampling." *Statistics and Computing*, 6: 113–119. 24

Woodard, D. B., Schmidler, S. C., and Huber, M. (2009). "Sufficient conditions for torpid mixing of parallel and simulated tempering." *Electronic Journal of Probability*, 14: 780–804.