2012

On thinning of chains in MCMC

William A. Link

Mitchell J. Eaton

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 112–115 doi: 10.1111/j.2041-210X.2011.00131.x

FORUM

On thinning of chains in MCMC

William A. Link and Mitchell J. Eaton

USGS Patuxent Wildlife Research Center, Laurel, MD 20708, USA

Summary

1. Markov chain Monte Carlo (MCMC) is a simulation technique that has revolutionised the analysis of ecological data, allowing the fitting of complex models in a Bayesian framework. Since 2001, there have been nearly 200 papers using MCMC in publications of the Ecological Society of America and the British Ecological Society, including more than 75 in the journal *Ecology* and 35 in the *Journal of Applied Ecology*.

2. We have noted that many authors routinely 'thin' their simulations, discarding all but every *k*th sampled value; of the studies we surveyed with details on MCMC implementation, 40% reported thinning.

3. Thinning is often unnecessary and always inefficient, reducing the precision with which features of the Markov chain are summarised. The inefficiency of thinning MCMC output has been known since the early 1990's, long before MCMC appeared in ecological publications.

4. We discuss the background and prevalence of thinning, illustrate its consequences, discuss circumstances when it might be regarded as a reasonable option and recommend against routine thinning of chains unless necessitated by computer memory limitations.

Key words: Markov chain Monte Carlo, thinning, WinBUGS

Introduction

Markov chain Monte Carlo (MCMC) is a technique (or more correctly, a family of techniques) for sampling probability distributions. Typical applications are in Bayesian modelling, the target distributions being posterior distributions of unknown parameters, or predictive distributions for unobserved phenomena. MCMC is becoming commonplace as a tool for fitting ecological models. The first applications of MCMC methods in publications of American and British ecological societies were in a paper published by the British Ecological Society (BES) in 2001 (Groombridge *et al.* 2001) and in five papers published by the Ecological Society of America (ESA) in 2002 (Gross, Craig, & Hutchison 2002; Link & Sauer 2002; Mac Nally & Fleishman 2002; O'Hara *et al.* 2002; Sauer & Link 2002). Since then, the use of MCMC in journals of these societies has increased rapidly. Summarising over three publications of the Ecological Society of America (*ESA*: *Ecology*, *Ecological Applications* and *Ecological Monographs*) and five publications of the British Ecological Society (*BES: J. of Ecology, J. of Applied Ecology*, *Functional Ecology*, *J. of Animal Ecology* and *Methods in Ecology and Evolution*), the numbers of publications using MCMC were 1, 6, 12, 10, 14, 21, 13, 28, 49 and 45, for years 2001–2010.

The appeal of MCMC is that it is almost always relatively easy to implement, even when the target distributions are complicated and conventional simulation techniques are impossible. The difference between MCMC and traditional simulation methods is that MCMC produces a dependent sequence – a Markov chain – of values, rather than a sequence of independent draws. The Markov chain sample is summarised just like a conventional independent sample; sample features (e.g. mean, variance and percentiles) are used to approximate corresponding features of the target distribution. The disadvantage of MCMC is that these approximations are typically less precise than would be obtained from an independent sample of the same size.

Many practitioners routinely thin their chains – that is, they discard all but every *k*th observation – with the goal of reducing autocorrelation. Among 76 *Ecology* papers published between 2002 and 2010, 15 mentioned MCMC, but did not apply it; eight used MCMC, but provided no details on the actual implementation. Twenty-one of the remaining 53 (40%) reported thinning; among these, the median rate of thinning was to select every 40th value ('×40' thinning). Five studies reported thinning rates of ×750 or higher, and the highest rate was $\times 10^5$. Among 73 papers published in five journals of the

*Correspondence author. E-mail: wlink@usgs.gov
Correspondence site: http://www.respond2articles.com/MEE/

BES, 27 mentioned MCMC but either did not apply it or used packaged software developed for genetic analyses that offered limited user-control over the implementation of MCMC. A further nine publications applied MCMC methods but provided no details on its implementation. Fifteen of the remaining 37 (41%) reported thinning of chains. The median thinning rate among these studies was ×29, and the highest was ×1000.

Our purpose in writing this note is to discourage the practice of thinning, which is usually unnecessary, and always inefficient. Our observation is not a new one: MacEachern & Berliner (1994) provide 'a justification of the ban [on] subsampling' MCMC output; see also Geyer (1992). We are not suggesting or promoting a ban on the practice; there are circumstances (discussed later) where thinning is reasonable. In these cases, we encourage the practitioner to be explicit in his or her reasoning for sacrificing one sort of efficiency for another. However, for approximation of simple features of the target distribution (e.g. means, variances and percentiles), thinning is neither necessary nor desirable; results based on unthinned chains are more precise.

We write this note assuming readers have some acquaintance with MCMC methods; for more details on fundamentals, we refer readers to Link *et al.* (2002) or to texts by Gelman *et al.* (2004) and Link & Barker (2010). Because our emphasis is on the practice of thinning chains, we assume that MCMC output follows from appropriate starting values and adequate burnin to allow evaluation as stationary chains.

Methods

We illustrate the counter-productive effects of thinning with two examples. The first is a simulation study of the relative performance of a specific Markov chain sampler; the second makes use of theoretical results for a two-state Markov chain, such as encountered in Bayesian multimodel inference.

EXAMPLE 1

Panel 1 describes a Markov chain produced by the Metropolis–Hastings algorithm. This particular chain produces samples from a t-distribution with m degrees of freedom. One begins by choosing a value $A > 0$; any value will do, though some will produce better chains than others, hence A is described as a 'tuning parameter'. Each step of the algorithm requires the generation of a pair (U_1, U_2) of random variables uniformly distributed on the interval [0,1] and a few simple calculations.

Consider the performance of this algorithm in drawing samples from the t-distribution with five degrees of freedom; our discussion focuses on chains produced using $A = 1$ or $A = 6$. History plots (X_t vs. t) are given for the first 1000 values of two chains in Fig. 1. Inspection of the graphs shows that the chain with $A = 6$ has a lower acceptance rate Pr $(X_t = X_*)$ than the chain with $A = 1$; the actual rates were 81·5% and 30·6% for $A = 1$ and $A = 6$, respectively.[1] Thus, the chain with $A = 1$ moves frequently, taking many small steps. A chain with $A = 50$ (not shown) has an acceptance rate of only 3·8%; it moves rarely and takes larger steps. Both extremes (A too small or too large) lead to poor MCMC performance, because consecutively sampled values are highly autocorrelated.

Panel 1. Metropolis–Hasting Markov chain algorithm for t-distribution with m degrees of freedom

Set $X_0 = 0$. Then, for $t = 1, 2, \ldots$

1. Generate $U_1, U_2 \sim U(0, 1)$
2. Set $X_* = X_{t-1} + A(2U_1 - 1)$
3. Calculate
$$r = \left(\frac{m+X_{t-1}^2}{m+X_*^2}\right)^{(m+1)/2}$$
4. If $U_2 < r$, set $X_t = X_*$. Otherwise, set $X_t = X_{t-1}$

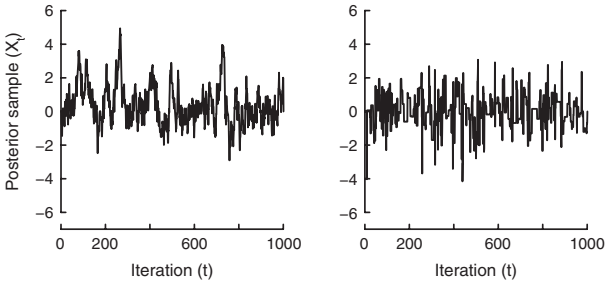

Fig. 1. History plots of chains of length 1000 from a Metropolis–Hastings sampler with tuning parameter $A = 1$ (left) and $A = 6$ (right).

Plots of the autocorrelation function (ACF) $f(h) = \rho(X_{t+h}, X_t)$ for the two chains are given in Fig. 2. Given a choice between the two, we would choose the chain with $A = 6$, because its sample values are more nearly independent. In practice, most users of MCMC rely on software like *WinBUGS* (Spiegelhalter *et al.* 2003) and are not directly involved in tuning the algorithms. *WinBUGS* does an admirable job of tuning its sampling, but with complex models, an ACF like that for the chain based on $A = 1$ is often the best that can be hoped for, or even better.

Note that the ACF for the chain with $A = 6$ is nearly zero at lag 10. We might thin the chain, taking every 10th observation and regarding these as independent. To achieve a comparable level of independence, we would need to take every 100th observation from a

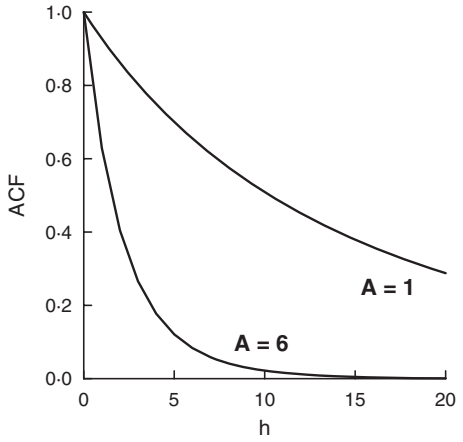

Fig. 2. Autocorrelation functions depicting the strength of the correlation between X_t and X_{t+h} (i.e. autocorrelation at lag h) for chains with $A = 1$ and $A = 6$.

[1]This and subsequent descriptions of the chains' performance are based on the average of results for 25 chains of length 250 000, and are accurate to the number of decimal places reported.

chain with $A = 1$. We wind up with a much smaller sample, but with less autocorrelation. The question is whether it is worth doing so.

We thus compare four MCMC sampling procedures: (1) with $A = 6$, unthinned; (2) with $A = 6$, thinning ×10; (3) with $A = 1$, unthinned; and (4) with $A = 1$, thinning ×100. We implemented each procedure for chains of length 10^4, 10^5 and 10^6 (before thinning). Each chain was summarised by its mean, standard deviation, 1st, 2·5th, 5th, 10th and 50th percentiles and replicated 1000 times.

For all of these parameters, summaries based on the unthinned chains tended to provide better estimates than those based on corresponding thinned chains (Tables 1 and 2). For example, consider estimates of the mean μ based on chains of length 10^6, with $A = 1$. In only 335 of 1000 replicate chains was the value based on the thinned chain closer to the true value than that from the unthinned chain (Table 1); the standard deviations among the approximations were 0·0134 and 0·0083, respectively, indicating a variance ratio (relative efficiency) of 2·6 in favour of using the unthinned chain (Table 2).

EXAMPLE 2

The Bayesian paradigm provides an appealing framework for inference in the presence of model uncertainty (Link & Barker 2006). The tasks of model selection (choosing a best supported model from a model set) and model weighting (combining inference across a collection of models with regard to their relative support by data) are dealt with in terms of probabilities on models in a model set. The mathematical formalism for model uncertainty involves cell probabilities for a latent categorical random variable M taking values in a s-dimensional state space $\mathcal{M} = (M_1, M_2, \ldots, M_s)$, (Link & Barker 2006). Here, the values M_j are models, and $\mathcal{M}$ is the model set. As in all Bayesian inference, prior probabilities for M are informed by data, and conclusions are based on posterior probabilities, $\eta_j = \Pr(M = M_j|Data)$. MCMC for M produces a Markov chain on M; the frequency with which this chain visits state M_j is used to estimate η_j.

Suppose that we are considering a two-model state space, that $\{X_t\}$ is a Markov chain of indicator variables for $M = M_1$, and that the process $\{X_t\}$ mixes slowly. Slow mixing means that transitions from $M = M_1$ to $M = M_2$ and vice versa are relatively infrequent, leading to high autocorrelation in the chain and reduced efficiency in estimating $\eta = \eta_1$.

For this simple Markov chain, it is possible to analytically evaluate the effect of autocorrelation on MCMC performance and to evaluate the 'benefit' (or otherwise) of thinning. Letting $\hat{\eta}$ denote the frequency with which $M = M_1$ and assuming an adequate burnin, $\hat{\eta}$ is unbiased for η and (to a very close approximation)

$$\mathrm{Var}(\hat{\eta}) = \frac{\eta(1-\eta)}{N} \times \frac{1+\theta}{1-\theta},$$

where N is chain length and θ is the lag one autocorrelation of the chain (see Appendix S1 for details on this formula and subsequent calculations).

It can be shown that taking every kth observation produces a chain with $N' = N/k$, $\eta' = \eta$ and $\theta' = \theta^k$. The ratio of variances for sample means (thinned chain relative to unthinned) is therefore

$$k \frac{1+\theta^k}{1-\theta^k} \times \frac{1-\theta}{1+\theta}, \qquad \text{eqn 1}$$

which is always >1: there is always a loss of efficiency because of thinning.

We recently used Bayesian multimodel inference to compare von Bertalanffy and logistic growth models for dwarf crocodiles (Eaton & Link 2011). We approximated posterior model probabilities using MCMC, producing a Markov chain of model indicators of length $N = 5\,000\,000$, with lag one autocorrelation $\theta = 0{\cdot}981$. Had we chosen to thin the chain by subsampling every 100th observation, the lag one autocorrelation would have been reduced to 0·151, but the chain length would have been reduced to 50,000; using eqn (1), we find that the variance of $\hat{\eta}$ would have increased by 28%.

Table 1. Probability that MCMC approximation based on thinned chain is closer to true value than approximation based on unthinned chain. Probabilities were estimated for mean ($\mu = 0$), standard deviation ($\sigma = \sqrt{5/3}$) and various percentiles $t_5(\alpha)$, for chains with $A = 6$ and $A = 1$, with unthinned chain lengths (UC Length) 10^4, 10^5 and 10^6. Probabilities were estimated based on 1000 replicate chains and are within ±0·03 of true values (95% CI)

A	UC length	μ	σ	t_5(0·01)	t_5(0·025)	t_5(0·05)	t_5(0·10)	t_5(0·50)
1	10^4	0·32	0·32	0·26	0·25	0·28	0·23	0·23
	10^5	0·31	0·37	0·30	0·29	0·25	0·24	0·22
	10^6	0·33	0·39	0·30	0·27	0·27	0·23	0·23
6	10^4	0·35	0·36	0·31	0·32	0·34	0·33	0·38
	10^5	0·32	0·40	0·30	0·32	0·33	0·34	0·35
	10^6	0·35	0·39	0·34	0·31	0·33	0·35	0·34

Table 2. Ratio of thinned chain variance vs. unthinned chain variance, among 1000 replicates. Ratios were calculated for mean (μ = 0), standard deviation ($\sigma = \sqrt{5/3}$) and various percentiles t_5 (α), for chains with $A = 1$ and $A = 6$, with unthinned chain lengths (UC Length) 10^4, 10^5 and 10^6

A	UC length	μ	σ	t_5(0·01)	t_5(0·025)	t_5(0·05)	t_5(0·10)	t_5(0·50)
1	10^4	2·7	1·8	4·2	3·7	4·2	5·1	6·7
	10^5	2·4	1·2	3·1	3·8	4·3	5·3	6·9
	10^6	2·6	1·3	3·1	3·7	4·5	5·4	6·8
6	10^4	1·9	1·1	2·2	2·3	2·4	2·2	1·7
	10^5	2·2	1·3	2·5	2·5	2·4	2·2	1·9
	10^6	2·1	1·1	2·5	2·6	2·6	2·2	1·8

Discussion

The greater precision associated with approximation from unthinned chains is not an artefact of the present examples, but an inevitable feature of MCMC (MacEachern & Berliner 1994). Indeed, this is not a surprising result; if one is interested in precision of estimates, why throw away data?

There are, in fact, several legitimate reasons for thinning chains. First, with independent samples, one can often estimate the precision of an MCMC approximation. So, in Example 1, one might apply ×10 thinning to a chain with $A = 6$, reducing a sample of size 10^6 to size 10^5, treating the resulting sample as independent random samples, and calculating $s/\sqrt{10^5}$ as a standard error. We did not see this offered as a motivation for thinning in any of the papers we reviewed but would suggest that even if it were, it would be better to report the mean of the unthinned chain as the estimate, and to use the standard error of the thinned chain as a conservative measure of precision. A better course of action, however, is to generate multiple independent chains [as, for example, when implementing the Gelman-Rubin diagnostic (Brooks & Gelman 1998)] to compute desired approximations for each chain, and to consider the variation among these independent values.

The reality is that too little attention is paid to the precision of MCMC approximations. We noted in our review of the 76 *Ecology* papers and 73 BES papers using MCMC that analysts often report 3 or 4 decimal place precision. This is rarely justified (Flegal, Haran, & Jones 2008). In Example 1, approximations based on unthinned $A = 6$ chains of length 10^6 have standard deviation of 0·0083; the third decimal place of the approximation is practically irrelevant. Even with an independent sample of size 10^6, the precision of the mean sample from the t_5 distribution is $\sqrt{5/3}/1000 = 0{\cdot}0013$. Many of the *Ecology* and BES papers had final sample sizes of 10 000 or less.

Another reason for thinning chains is (or used to be) limitations in computer memory and storage. High autocorrelation might be unavoidable, requiring very long chains. With many nodes monitored, memory and storage limitations can be a consideration. It is often possible to circumvent these limitations without too much difficulty, but the time spent in programming such a solution might not be worth the trouble, making thinning an inviting option.

Finally, it might make sense to thin chains if a great deal of post-processing is required. It may be that a derived parameter must be calculated for each sampled value of the Markov chain. The derived parameter might be the result of complex matrix calculations, or even the result of a simulation – e.g., from a population viability analysis. Given that these calculations impose a substantial computational burden, overall results might be improved by paying greater attention to reduce autocorrelation in the chains being used.

Our point in writing this note is not to suggest that the practice of thinning MCMC chains is never appropriate, and thus should be banned, but to highlight that there is nothing advantageous or necessary in it *per se*. In most cases, greater precision is available by working with unthinned chains.

Acknowledgements

We thank JR Sauer, JA Royle, Marc Kéry and one anonymous reviewer for helpful comments and discussion in the preparation of this manuscript. Use of trade, product or firm names does not imply endorsement by the US Government. Use of trade, product or firm names does not imply endorsement by the US Government.

References

Brooks, S.P. & Gelman, A. (1998) Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Eaton, M.J. & Link, W.A. (2011) Estimating age from recapture data: integrating incremental growth measures with ancillary data to infer age-at-length. *Ecological Applications*. in press, doi: 10.1890/10-0626.1.

Flegal, J.M., Haran, M. & Jones, G.L. (2008) Markov chain Monte Carlo: can we trust the third significant figure? *Statistical Science*, **23**, 250–260.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2004) *Bayesian Data Analysis*, 2nd edn. Chapman and Hall, New York.

Geyer, C.J. (1992) Practical Markov Chain Monte Carlo. *Statistical Science*, **7**, 473–483.

Groombridge, J.J., Bruford, M.W., Jones, C.G. & Nichols, R.A. (2001) Evaluating the severity of the population bottleneck in the Mauritius kestrel Falco punctatus from ringing records using MCMC estimation. *Journal of Animal Ecology*, **70**, 401–409.

Gross, K., Craig, B.A. & Hutchison, W.D. (2002) Bayesian estimation of a demographic matrix model from stage-frequency data. *Ecology*, **83**, 3285–3298.

Link, W.A. & Barker, R.J. (2006) Model weights and the foundations of multimodel inference. *Ecology*, **87**, 2626–2635.

Link, W.A. & Barker, R.J. (2010) *Bayesian Inference: With Ecological Applications*. Elsevier/Academic Press, Amsterdam.

Link, W.A. & Sauer, J.R. (2002) A hierarchical analysis of population change with application to Cerulean Warblers. *Ecology*, **83**, 2832–2840.

Link, W.A., Cam, E., Nichols, J.D. & Cooch, E. (2002) Of BUGS and birds: Markov chain Monte Carlo for hierarchical modeling in wildlife research. *The Journal of Wildlife Management*, **66**, 277–291.

Mac Nally, R. & Fleishman, E. (2002) Using "Indicator" species to model species richness: model development and predictions. *Ecological Applications*, **12**, 79–92.

MacEachern, S.N. & Berliner, L.M. (1994) Subsampling the Gibbs sampler. *The American Statistician*, **48**, 188–190.

O'Hara, R.B., Arjas, E., Toivonen, H. & Hanski, I. (2002) Bayesian analysis of metapopulation data. *Ecology*, **83**, 2408–2415.

Sauer, J.R. & Link, W.A. (2002) Hierarchical modeling of population stability and species group attributes from survey data. *Ecology*, **83**, 1743–1751.

Spiegelhalter, D.J., Thomas, A., Best, N.G. & Lunn, D. (2003) *WinBUGS User Manual. Version 1.4*. Medical Research Council Biostatistics Unit, Cambridge, UK.

Received 22 February 2011; accepted 18 May 2011
Handling Editor: David Warton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Derivation of variance formula for sample state frequency of a two-state Markov chain. This formula is used to demonstrate the loss of precision resulting from thinning of chains; the variance associated with a thinned chain is always larger than that associated with the original unthinned chain.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.