Simulated Tempering for the MCMC Package

Charles J. Geyer

December 17, 2011

1 Parallel and Serial Tempering

Serial tempering (Marinari and Parisi, 1992; Geyer and Thompson, 1995) runs a Markov chain whose state is (i, x), where i is a positive integer between 1 and k and x is an element of $\mathbb{R}^p$. The unnormalized density of the equilibrium distribution is $h(i, x)$. The integer i is called the *index of the component of the mixture*, and the integer k is called the *number of components of the mixture*. The reason for this terminology is that

$$h(x) = \sum_{i=1}^{k} h(i, x), \tag{1}$$

which is the unnormalized marginal density of x derived from the unnormalized joint density $h(i, x)$ of the equilibrium distribution of the Markov chain, is a mixture of k component distributions having unnormalized density $h(i, \cdot\,)$ for different i.

Parallel tempering (Geyer, 1991) runs a Markov chain whose state is $(x_1, \ldots, x_k)$ where each x_i is an element of $\mathbb{R}^p$. Thus the state is a vector whose elements are vectors, which may be thought of as a $k \times p$ matrix. The unnormalized density of the equilibrium distribution is

$$h(x_1, \ldots, x_k) = \prod_{i \in I} h(i, x_i). \tag{2}$$

This joint equilibrium distribution is the product of the marginals $h(i, \cdot\,)$ for different i. This the x_i are asymptotically independent in parallel tempering.

2 Sensitivity to Normalization

So long as one is only interested in one of the component distributions $h(i, \cdot\,)$, both parallel and serial tempering do the job. And this job is

what gives them the name "tempering" by analogy with simulated annealing (Marinari and Parisi, 1992). The other component distributions only help in sampling the component of interest. In this job, parallel tempering is easier to set up because it is insensitive to normalizing constants in the following sense. Suppose we change the normalization for each component distribution using

$$h^*(i, x) = a_i h(i, x).$$

This greatly changes the mixture distribution (1) sampled by simulated tempering. We now get

$$h^*(x) = \sum_{i=1}^{k} h^*(i, x) = \sum_{i=1}^{k} a_i h(i, x),$$

which may be very different from (1), even considered as an unnormalized density (which it is). But (2), considered as an unnormalized density (which it is), does not change at all

$$\begin{aligned} h^*(x_1, \ldots, x_k) &= \prod_{i=1}^{k} h^*(i, x) \\ &= \prod_{i=1}^{k} a_i h(i, x) \\ &= \left(\prod_{i=1}^{k} a_i \right) \left(\prod_{i=1}^{k} h(i, x) \right) \\ &= \left(\prod_{i=1}^{k} a_i \right) h(x_1, \ldots, x_k) \end{aligned}$$

(the normalizing constant changes, but that does not matter for an unnormalized density; it still specifies the same probability distribution). All this is to say that serial tempering is very sensitive to the choices of normalizing constants of the individual component distributions (the a_i in the preceding discussion) and parallel tempering is totally insensitive to them. Thus parallel tempering is easier to set up and get working. Geyer and Thompson (1995), however, independently invented serial tempering because it worked for a problem where parallel tempering failed. So for this "tempering" job, where one is only interested in sampling one component distribution (and the others are just helpers) parallel tempering is easier to use but serial tempering works better.

3 Umbrella Sampling

Sometimes one is actually interested in sampling a particular mixture distribution having unnormalized density (1). This arises in Bayesian and frequentist model averaging and for other reasons. An umbrella term for this application is "umbrella sampling" (Torrie and Valleau, 1977). In this application only serial tempering does more than parallel tempering. Parallel tempering can simulate any directly specified mixture. If

$$f(i, x) = \frac{h(i, x)}{\int h(i, x)\, dx}$$

are the normalized component distributions and $b_1, \ldots, b_k$ are nonnegative and sum to one, then

$$f(x) = \sum_{i=1}^{k} b_i f(i, x)$$

is a normalized mixture distribution, and parallel tempering can be used to sample it. However, this "directly specified" mixture is often not of interest because the individual component normalizing constants

$$c_i = \int h(i, x)\, dx \tag{3}$$

are unknown. Suppose we are doing Bayesian model averaging and $h(i, x)$ is the unnormalized posterior density (likelihood times prior). This means i and x are parameters to the Bayesian; i denotes the model and x denotes the within-model parameters. Then the normalizing constants (3) are unnormalized Bayes factors, which Bayesians use for model comparison.

The function $i \mapsto c_i$ is the unnormalized density of the marginal distribution of the random variable i derived from the joint distribution $h(i, x)$, which is the equilibrium distribution of the Markov chain. It is therefore estimated, up to a constant of proportionality, by the marginal distribution of i. Thus serial tempering, unlike parallel tempering, provides simple and direct estimates of Bayes factors and other normalizing constants.

4 Update Mechanisms

Traditionally, tempering makes two kinds of elementary updates, one changes only x in serial tempering or one x_i in parallel tempering. We call them within-component updates, and will use normal random walk Metropolis updates analogous to those used by the `metrop` function. The other kind

changes i in serial tempering or swaps x_i and x_j in parallel tempering. We call them jump/swap updates (jump in serial tempering, swap in parallel tempering).

4.1 Serial Tempering

The combined update is a 50-50 mixture of within-component elementary updates and jump updates. Suppose the current state is (i, x). A within-component update proposes x^* which is normally distributed centered at x. Then Metropolis rejection of the proposal is done with Metropolis ratio

$$\frac{h(i, x^*)}{h(i, x)}$$

This is valid because the proposal is symmetric by symmetry of the normal distribution. A jump update proposes i^*, which is chosen uniformly at random from the "neighbors" of i (the neighbor relation is specified by a user-supplied logical matrix). This proposal need not be symmetric, because i and i^* need not have the same number of neighbors. Write $n(i)$ and $n(i^*)$ for these neighbor counts. Then the probability of proposing i^* when the current state is i is $1/n(i)$, and the probability of proposing i when the current state is i^* is $1/n(i^*)$. Hence the appropriate Hastings ratio for Metropolis-Hastings rejection is

$$\frac{h(i^*, x)/n(i^*)}{h(i, x)/n(i)} = \frac{h(i^*, x)}{h(i, x)} \cdot \frac{n(i)}{n(i^*)}$$

4.2 Parallel Tempering

The combined update is a 50-50 mixture of within-component elementary updates and swap updates. Suppose the current state is $(x_1, \ldots, x_k)$. A within-component chooses i uniformly at random in $\{1, \ldots, k\}$, and then proposes x_i^* which is normally distributed centered at x_i. Then Metropolis rejection of the proposal is done with Metropolis ratio

$$\frac{h(i, x_i^*)}{h(i, x_i)}$$

This is valid because the proposal is symmetric by symmetry of the normal distribution. A swap update chooses i uniformly at random in $\{1, \ldots, k\}$ and then j, which is chosen uniformly at random from the neighbors of i. This proposal is automatically symmetric, because a swap move is its

own inverse. Hence the appropriate Hastings ratio for Metropolis-Hastings rejection is

$$\frac{h(i, x_j)h(j, x_i)}{h(i, x_i)h(j, x_j)}$$

5 Acceptance Rates

Metropolis-Hastings acceptance rates are not comparable to Metropolis acceptance rates. For serial tempering

$$E\left\{1 \wedge \frac{h(i^*, x)}{h(i, x)} \cdot \frac{n(i)}{n(i^*)}\right\} \neq E\left\{1 \wedge \frac{h(i^*, x)}{h(i, x)}\right\}$$

where the expectations are taken with respect to (i, x) having the equilibrium distribution of the Markov chain and the conditional distribution of i^* given i being uniform over neighbors of i. For parallel tempering,

$$E\left\{1 \wedge \frac{h(i, x_j)h(j, x_i)}{h(i, x_i)h(j, x_j)} \cdot \frac{n(i)}{n(j)}\right\} \neq E\left\{1 \wedge \frac{h(i, x_j)h(j, x_i)}{h(i, x_i)h(j, x_j)}\right\}$$

where the expectations are taken with respect to $(x_1, \ldots, x_k)$ having the equilibrium distribution of the Markov chain, (i, j) being independent of $(x_1, \ldots, x_k)$, the marginal distribution of i being uniform on $\{1, \ldots, k\}$, and the conditional distribution of j given i being uniform over neighbors of i.

Thus we need to report rates going both ways, for example, for serial tempering, when $i = 1$ and $i^* = 2$ as well as when $i = 2$ and $i^* = 1$. And similarly for parallel tempering.

References

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proceedings of the Symposium on Interface Critical Applications of Scientific Computing (23rd): Biology, Engineering, Medicine, Speech Held in Seattle, Washington on 21-24 April 1991*, J. R. Kettenring and E. M. Keramidas, eds., 156–163. `http://www.stat.umn.edu/geyer/f05/8931/c.ps`

Geyer, C. J., and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, **90**, 909–920.

Marinari, E., and Parisi G. (1992). Simulated tempering: A new Monte Carlo Scheme. *Europhysics Letters*, **19**, 451–458.

Torrie, G. M., and Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, **23**, 187–199.