Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference

Charles J. Geyer; Elizabeth A. Thompson

Journal of the American Statistical Association, Vol. 90, No. 431 (Sep., 1995), 909-920.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199509%2990%3A431%3C909%3AAMCMCW%3E2.0.CO%3B2-R

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Thu Nov 3 18:27:38 2005

Annealing Markov Chain Monte Carlo With Applications to Ancestral Inference

Charles J. GEYER and Elizabeth A. THOMPSON*

Markov chain Monte Carlo (MCMC; the Metropolis–Hastings algorithm) has been used for many statistical problems, including Bayesian inference, likelihood inference, and tests of significance. Though the method generally works well, doubts about convergence often remain. Here we propose MCMC methods distantly related to simulated annealing. Our samplers mix rapidly enough to be usable for problems in which other methods would require eons of computing time. They simulate realizations from a sequence of distributions, allowing the distribution being simulated to vary randomly over time. If the sequence of distributions is well chosen, then the sampler will mix well and produce accurate answers for all the distributions. Even when there is only one distribution of interest, these annealing-like samplers may be the only known way to get a rapidly mixing sampler. These methods are essential for attacking very hard problems, which arise in areas such as statistical genetics. We illustrate the methods with an application that is much harder than any problem previously done by MCMC, involving ancestral inference on a very large genealogy (7 generations, 2,024 individuals). The problem is to find, conditional on data on living individuals, the probabilities of each individual having been a carrier of cystic fibrosis. Exact calculation of these conditional probabilities is infeasible. Moreover, a Gibbs sampler for the problem would not mix in a reasonable time, even on the fastest imaginable computers. Our annealing-like samplers have mixing times of a few hours. We also give examples of samplers for the "witch's hat" distribution and the conditional Strauss process.

KEY WORDS: Cystic fibrosis; Gibbs sampler; Metropolis algorithm; Pedigree analysis; Simulated tempering; Stochastic approximation.

1. INTRODUCTION

Markov chain Monte Carlo (MCMC) in the form of the Metropolis–Hastings algorithm (Hastings 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953) and its special case the Gibbs sampler (Geman and Geman 1984) has been used in recent years to attack a wide variety of intractable statistical problems. (See, for example, Besag and Green 1993, Geyer 1992, Geyer and Thompson 1992, Smith and Roberts 1993, Tierney 1994, and the accompanying discussions and references.) MCMC simulates realizations from probability distributions whose densities are known up to a normalizing factor. If $h(x)$ is a nonnegative integrable function on the sample space, then the Metropolis–Hastings algorithm simulates a Markov chain whose equilibrium distribution is proportional to $h(x)$ using only evaluations of $h(x)$.

If the chain is irreducible, then time averages over the chain converge to expectations with respect to the stationary distribution as the Monte Carlo sample size goes to infinity; but if the chain is slowly mixing, then it may take astronomically large sample sizes to get accurate estimates. Slow mixing typically occurs in problems where the sample space has high dimension. For samplers that update one variable at a time, like the Gibbs sampler, the mixing time can be exponential in the number of variables. Thus, to do MCMC on high-dimensional problems, it is necessary to make a radical change in the sampling scheme, getting away from updating one variable at a time. The first such method was the Swendsen–Wang (1987) algorithm for the Ising model and related models of statistical physics. A number of similar algorithms have been devised since (Besag and Green 1993; Wang and Swendsen 1990) and are grouped under the name "cluster algorithms." Although these algorithms are highly effective, they seem to apply only to problems where all variables are conditionally positively correlated given the rest.

A much more general algorithm was proposed by Geyer (1991a) under the name "Metropolis-coupled MCMC" (MCMCMC). An improvement of MCMCMC by changing from parallel simulation of distributions at different temperatures to random temperatures led us to the algorithm that we called "pseudo-Bayes" in the first version of this article. We later found that the key idea had been independently proposed by Marinari and Parisi (1992) under the name "simulated tempering." We have adopted their term even though our algorithm differs from theirs in some details and adds a number of ideas needed to make it work on a wide variety of problems. This article explains our version of simulated tempering and provides examples of its use.

Both MCMCMC and simulated tempering are based on an analogy with simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983). Simulated annealing is an algorithm for optimization rather than Monte Carlo, but it provides the useful metaphor of starting with "heated" versions of the problem and slowly cooling down to the problem of interest. Because an MCMC algorithm is a Markov chain with stationary transition probabilities, neither MCMCMC or simulated tempering "cools" like simulated annealing, but both use a one-parameter family of probability distributions indexed by a parameter called "temperature," ranging from

* Charles J. Geyer is Assistant Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Elizabeth A. Thompson is Professor, Department of Statistics, University of Washington, Seattle, WA 98195. The research of Geyer was supported in part by National Science Foundation (NSF) Grant DMS-9007833; that of Thompson by NSF Grant BIR-9305835 and National Institutes of Health Grant GM-46255. The authors thank K. Morgan for access to data on the Hutterite genealogy. These data were compiled by T. M. Fujiwara, K. Morgan, and J. Crumley with support from the Canadian Genetic Diseases Network. The authors thank Augie Kong for discussions about MCMC—in particular, his explanation that Bayesian "data augmentation" versions of Gibbs sampling are MCMCMC in time instead of space, which is suggestive of simulated tempering. They also thank Peter Green, Neal Madras, and the editor, associate editors, and referees for helpful comments and for pointing out various references, particularly Peter Green for Marinari and Parisi (1992).

© 1995 American Statistical Association

the distribution of interest as the "coldest" temperature to a "hottest" distribution that is much easier to simulate.

There have been other proposals in the statistics literature for speeding up the mixing of MCMC samplers, such as using the classical variance reduction methods of ordinary independent-sample Monte Carlo, like importance sampling and antithetic variates (see Besag and Green 1993 and Tierney, in press, and the references cited therein), but those methods only reduce the mixing time by a constant factor and would not change the exponential growth of mixing time with dimension. There have also been other proposals in the statistical physics literature. Berg and Neuhaus (1991), following earlier work by Berg and other authors, proposed simulating a "multicanonical ensemble" as the stationary distribution of the sampler and reweighting the multicanonical ensemble to the distribution of interest by the importance sampling formula. This is similar to work by Torrie and Valleau (1977), who called their importance sampling scheme "umbrella sampling," except that Torrie and Valleau did not present their method as a way of doing intractable high-dimensional problems but rather as one for obtaining stable estimates of expectations with respect to a wide range of distributions in the same spirit as the method of "reweighting mixtures" of Geyer (1991b). Frantz, Freeman, and Doll (1990) proposed a method called "J-walking" that is not an exact MCMC scheme, because it does not run a Markov chain with a specified stationary distribution, but rather only an approximation thereof. If it were corrected so as to be exact, then it would be MCMCMC.

We provide three examples of our simulated tempering method. The "witch's hat" distribution provides an illustration of how simulated tempering works when Gibbs sampling fails. A more realistic example is the Strauss process, where the method is used for importance sampling in the spirit of Torrie and Valleau (1977) and Geyer (1991b). The third example is from pedigree analysis. We analyze large 2,024-member and 5,277-member pedigrees for which simulated tempering seems to be the only known feasible sampling algorithm. Although our methods were developed to do high-dimensional problems like those in pedigree analysis, they can be applied to any situation in which MCMC is used. In easier problems, these annealing-like samplers go a long way toward alleviating concerns about convergence.

2. ALGORITHMS

Both MCMCMC and simulated tempering simulate a sequence of m distributions specified by unnormalized densities $h_i(x)$, $i = 1, \ldots, m$ on the same sample space, where the index i is called "temperature." We call $h_1(x)$ the "cold" distribution and $h_m(x)$ the "hot" distribution. Sometimes, as in Section 4, all m distributions are of interest, but usually only the cold distribution is of interest, and the rest are used only to increase the mixing.

Simulated annealing uses a specific form of "heating" a problem that is sometimes called "powering up." If $h(x)$ is the unnormalized density for the distribution of interest, then $h(x)^{1/\beta}$ for $\beta > 1$ are the "heated" unnormalized densities, including perhaps $\beta = \infty$, which gives $h(x) \equiv 1$. This form comes from statistical physics, where the distribution of a thermodynamic equilibrium has an unnormalized density of the form $e^{-U(x)/kT}$, where $U(x)$ is the energy function of the system, T is the absolute temperature, and k is the Boltzmann constant. Such a distribution, called a Gibbs distribution, gives the Gibbs sampler its name. It is a special case of powering up with $h(x) = e^{-U(x)}$ and $\beta = kT$. Powering up heating is natural for a Gibbs distribution. Marinari and Parisi (1992) used it for their example, the random-field Ising model, and we use it for the conditional Strauss process example, both Gibbs distributions. But powering up is *not* an essential part of simulated tempering or MCMCMC. In the "witch's hat" example in Section 3, powering up is useless, but a different form of heating works well.

2.1 Simulated Tempering

For now, suppose that the $h_i(x)$ have been specified; guidance for choosing them is given later. Also suppose that there is available for each i a method for updating x that has $h_i(x)$ as a stationary distribution; a Gibbs or Metropolis update for $h_i(x)$, for example. The state of a simulated tempering sampler is the pair (x, i), where x takes values in the common state space of all the $h_i(x)$ and the temperature i is now random. The stationary distribution of the sampler is proportional to $h_i(x)\pi(i)$, where $\pi(1), \ldots, \pi(m)$ are auxiliary numbers that must be chosen in advance. We call π the *pseudoprior* because $h_i(x)\pi(i)$ looks like the product of likelihood and prior, i being the parameter and x the data, and because it determines the distribution of temperatures.

The specification of one iteration of the "Hastings version" of the simulated tempering algorithm is as follows:

1. Update x using a Metropolis–Hastings or Gibbs update for h_i.
2. Set $j = i \pm 1$ according to probabilities $q_{i,j}$, where $q_{1,2} = q_{m,m-1} = 1$ and $q_{i,i+1} = q_{i,i-1} = \frac{1}{2}$ if $1 < i < m$.
3. Calculate the Hastings ratio,

$$r = \frac{h_j(x)\pi(j)}{h_i(x)\pi(i)} \frac{q_{j,i}}{q_{i,j}},$$

and accept the transition (set i to j) or reject it according to the Metropolis rule: accept with probability $\min(r, 1)$.

In the calculation of r in Step 3, the factor $q_{j,i}/q_{i,j}$ is the Hastings (1970) modification of the Metropolis algorithm. It compensates for the asymmetric proposals. A "Metropolis version" of the algorithm uses the probabilities $q_{1,1} = q_{1,2} = q_{m,m-1} = q_{m,m} = \frac{1}{2}$ in Step 2, so the factor $q_{j,i}/q_{i,j}$ in Step 3 disappears. Because half the time it does not attempt to move from $i = 1$ or $i = m$, the Metropolis version makes fewer transitions and is slightly inferior.

There are two built-in diagnostics. First, any pair of adjacent distributions that are too far apart will be indicated by low acceptance rates in Step 3. Second, consider the *occupation numbers* of the chain, the number of iterations spent in each temperature i. If the sampler does not mix, then the occupation numbers will be very uneven. This indicates the need for a better pseudoprior. Simulated tempering has advantages over MCMCMC (Geyer 1991a), in that we keep only one copy of the state x rather than m copies, so the

chain uses less storage and also mixes better. The disadvantage is that simulated tempering needs a good pseudoprior, which must be determined by preliminary experimentation.

2.2 How Many Distributions?

The dynamics of a simulated tempering sampler are complex, so it is difficult to give criteria for choosing the number and spacing of the distributions, but some intuition can be obtained from examining a simplified model. Consider a random walk on the integers $1, \ldots, m$ having transitions to adjacent states with probability $p/2$ and staying at the same point with probability $1 - p$ for the interior points and $1 - p/2$ for the endpoints. This is a random walk with reflecting barriers at $x = \frac{1}{2}$ and $x = m + \frac{1}{2}$ (Feller 1968) and models a simulated tempering sampler with constant acceptance rate p independent of the state. Various properties could be called the mixing time of this random walk; here we consider the expected time taken to move from one end to the other. Using the methods of Feller (1968, chap. XIV), the expected time to go from $x = 1$ to $x = m$ is $m(m - 1)/p$.

This suggests that acceptance rates should not be too large. Certainly, it is a losing proposition to double the number of distributions unless doing so multiplies the acceptance rate by a factor of 4. When the acceptance rate is already above 25%, this is not possible. The actual sampler may behave rather differently from the random walk model, however, so we recommend acceptance rates in the range of 20 to 40%. This agrees with the behavior of our pedigree examples (Sec. 5.4). It is not always possible, though, to obtain acceptance rates this low (Sec. 3), no matter how wide the temperature gaps are. The problem is that acceptance rates averaged over the whole sample space may not be reflective of acceptance rates in parts of the sample space that are important for mixing (Sec. 3). Although average acceptance rates may not be a sufficient guide, we have no better proposal at this time.

2.3 Adjusting the Pseudoprior and the Spacing

The stationary distribution of a simulated tempering Markov chain is a joint distribution for the pair (X, I), where X is a random realization of the state variable x and I is a random realization of the "temperature" i. The marginal distribution of I is

$$\Pr(I = i) \propto \pi(i) \int h_i(x)\, d\mu(x) = c(i)\pi(i),$$

where $c(i) = \int h_i\, d\mu$ is the normalizing constant for distribution i. Hence, if $\pi(i) = 1/c(i)$, then the marginal distribution of I is uniform, the sampler spends $1/m$ of the time sampling each distribution, and there is no temperature that is not visited frequently. If the gaps between the distributions are also adjusted so that the acceptance rates for jumps between h_i and h_{i+1} are not too large, then there will be no bottlenecks and the sampler will rapidly move from any distribution to any other distribution. By the argument of the preceding section, we also do not want the gaps between distributions to be too small. We usually want an acceptance rate between 20% and 40%.

Suppose that we have decided on a form of heating by specifying a one-parameter family of unnormalized densities h_λ, $0 \le \lambda \le 1$, with h_0 as the cold distribution and h_1 as the hot distribution. The adjustment problem is to find a finite set of λ's satisfying $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_m = 1$ such that the simulated tempering sampler with distributions h_{λ_i} has uniform acceptance rates at a specified level when the pseudoprior is adjusted to be the inverse normalizing constants. Thus we have two adjustment problems to solve: adjusting the pseudoprior for fixed λ's, and adjusting the λ's. In principle one could try to solve both problems at once, but it is simpler and easier to consider these as two separate problems and to iterate between the two.

For adjusting the pseudoprior, we offer three methods: (a) an iterative adjustment method, (b) an MCMCMC sampler, and (c) stochastic approximation. For adjusting the λ's, we offer just one method.

a. If the pseudoprior is already well enough adjusted so that the sampler mixes, then we can estimate the normalizing constants empirically up to a constant of proportionality by

$$\hat{c}(i) = \frac{\widehat{\Pr}(I = i)}{\pi(i)}, \tag{1}$$

where $\widehat{\Pr}(I = i)$ is the fraction of time the sampler spends in the ith distribution, which is proportional to the occupation numbers $o(i)$. This serves as a method of estimating normalizing constants and also for adjusting pseudopriors. For the next run, $\pi(i)/o(i)$ is used as the pseudoprior.

b. If an MCMCMC sampler using the same sequence of distributions as the simulated tempering sampler mixes, then a preliminary run of the MCMCMC can be used to estimate the normalizing constants, either by direct Monte Carlo integration (Geyer and Thompson 1992) or by reverse logistic regression (Geyer 1991b).

c. Stochastic approximation, or the Robbins–Munro method (Wasan 1969), for simulated tempering starts with any values for the pseudoprior and updates the values as the chain progresses. At iteration k, the amount $c_0/[m(k + n_0)]$ is added to $\log \pi(i)$ for each i not equal to the current state I, and the amount $c_0/(k + n_0)$ is subtracted from $\log \pi(I)$. Here c_0 and n_0 are positive constants chosen by the user. It is necessary to choose a c_0 small enough and n_0 large enough so that the algorithm does not make large overcorrections early in the run; but if c_0 is too small or n_0 too large, then it will take too long for the algorithm to converge to a useful pseudoprior.

Consider now adjustment of the λ's. Suppose that the observed acceptance rates for a run were $a_1, \ldots, a_{m-1}$ for the $m - 1$ gaps between the m distributions. For $1 < i < m - 1$, we can take a_i to be the average of the rates for transitions going up and down between distributions i and $i + 1$. For a Hastings sampler, we take a_1 to be the acceptance rate going up plus half the rate going down, because transitions up are proposed twice as frequently and accepted half as often, and we correct a_{m-1} similarly.

We take as a model of the acceptance rate that the rate for transitions between h_{λ_i} and $h_{\lambda_{i+1}}$ is

$$a_i = \exp\left(-\int_{\lambda_i}^{\lambda_{i+1}} b(s)\,ds\right), \tag{2}$$

where $b(s)$ is some unknown function. We estimate $b(s)$ as a step function that is constant on the intervals between the λ_k,

$$b(s) = b_i = \frac{1}{\lambda_{i+1} - \lambda_i} \log \frac{1}{a_i}, \qquad \lambda_i < s < \lambda_j.$$

Then new intervals are determined with endpoints $\lambda_1^*, \lambda_2^*, \ldots$ that have a specified acceptance rate α according to the model (2).

Although the model is ad hoc, it works well as long as the adjustments are not too large. It does tend to overshoot in its corrections, however. If the observed acceptance rates are about 90% and one asks for 30%, it may produce λ^*'s that give acceptance rates varying from 10% to 30%. Another iteration, collecting another sample and making another adjustment, will give approximately uniform acceptance rates at the desired level.

In practice, these methods are all used in conjunction. We usually start with a few distributions at the hot end, running first with λ's equally and narrowly spaced. First, normalizing constants are determined by stochastic approximation during the run. Second, another run with stochastic approximation turned off checks that the first run converged, and the pseudoprior is adjusted using (1). Third, the spacing of the λ's is adjusted using the model (2) and the normalizing constants or the new λ's are determined by cubic spline interpolation. Fourth, a final run with stochastic approximation turned off checks that the new λ's do give the desired acceptance rates; if they do not, then the λ's are adjusted again. This completes the cycle for the current set of distributions. More distributions are then added, with the spacing of the λ's and the normalizing constants determined by extrapolation from those already determined. The whole cycle of adjustments is then repeated for the augmented distribution set.

For the hardest problems we have done, we added 5 distributions in each cycle, so 8 cycles were needed for 40 distributions; the last cycle takes almost half the running time. Having finally obtained a sampler that mixes and samples the cold distribution, we do one fairly long run and a final adjustment using (1) and (2). In our experience, a useful pseudoprior can be found in an amount of time that is roughly of the same order that spent running the sampler once the pseudoprior has been determined. Note that a simulated tempering sampler has the correct stationary distribution for any strictly positive pseudoprior. It will mix faster if the pseudoprior approximates the inverse normalizing constants, but high precision in the approximation is not necessary.

2.4 Regeneration

Some Markov chains can be made to regenerate, and this can improve estimation (Ripley 1987). This is easily done with simulated tempering. Choose the hot distribution $h_m(x)$ so that independent sampling is possible, and when $i = m$ in Step 1 of the algorithm, update x with an independent sample from h_m. Given $i = m$, the next value of x does not depend on the current value, and the future path of the chain is independent of the past. The set of states (x, i) such that $i = m$ (x arbitrary) is an atom of the Markov chain, times when $i = m$ are regeneration times, and segments of the sample path between regeneration times (called *tours*) are stochastically independent.

Regeneration greatly simplifies estimation of Monte Carlo error. It also eliminates "start up bias" if we start at the atom (at temperature m) and run until another regeneration time, so the sample path consists of a number of complete tours. Let τ_k, $k = 0, \ldots, K$, with $\tau_0 = 0$, be the regeneration times. The sample path is (X_t, I_t) for $t = 1, \ldots, \tau_K$, and $I_0 = m$. (The value of X_0 is irrelevant.) By an analog of Wald's lemma in sequential sampling (Nummelin 1984, pp. 76, 81) the expectation over a complete tour is unbiased:

$$E \sum_{t=\tau_{k-1}+1}^{\tau_k} g(X_t, I_t) = E(g(X, I))E(\tau_1),$$

where $Eg(X, I)$ is an expectation with respect to the stationary distribution and the other two expectations are with respect to the distribution of the Markov chain.

If we are trying to determine the expectation of $f(X)$ under the cold distribution $E(f(X) \mid I = 1)$, then we calculate the sums

$$Z_k = \sum_{t=\tau_{k-1}+1}^{\tau_k} f(X_t)w(I_t)$$

and

$$N_k = \sum_{t=\tau_{k-1}+1}^{\tau_k} w(I_t)$$

for $k = 1, \ldots, K$, where $w(I)$ is 1 when $I = 1$ and 0 otherwise. Then the Z_k are iid with expectation $E(f(X)w(I))E(\tau_1)$, and the N_k are iid with expectation $E(w(I))E(\tau_1)$. Hence by the ergodic theorem,

$$\frac{Z_1 + \cdots + Z_K}{N_1 + \cdots + N_K} \to \frac{E(f(X)w(I))}{E(w(I))} = E(f(X) \mid I = 1). \tag{3}$$

If the variances of Z_k and N_k can be shown to be finite, then the standard error of the Monte Carlo estimate can be calculated using the ratio estimator from finite population sampling (Ripley 1987, pp. 158 ff.). Let $\hat{\mu}_K$ denote the left side of (3) and μ denote the right side. Let $V_k = Z_k - \mu N_k$. Then the V_k are iid mean zero random variables with finite variance (say σ_V^2) that can be estimated by $\hat{\sigma}_V^2 = 1/K \sum_{k=1}^K V_k^2$. Now $K^{-1/2}(V_1 + \cdots + V_K)$ is asymptotically Normal$(0, \sigma_V^2)$, so

$$\sqrt{K}(\hat{\mu}_K - \mu) = \frac{\frac{1}{\sqrt{K}}(V_1 + \cdots + V_K)}{\frac{1}{K}(N_1 + \cdots + N_K)}$$

converges to Normal$(0, \sigma_V^2/\nu^2)$, where ν is the expectation of the N_k. Thus the asymptotic variance of $\hat{\mu}_K$ can be estimated by $(\hat{\sigma}_V^2/\hat{\nu}^2)/K$, where $\hat{\nu}$ is the sample mean of the N_k.

Typically, only a small fraction of tours will visit the cold distribution, so most of the N_k will be zero. One can instead average only over "informative tours" for which N_k is nonzero; one obtains the same mean and variance estimates either way, provided that K rather than $K - 1$ is used in computing $\hat{\sigma}^2_V$.

It is not necessary that the number of tours K be fixed in advance of the run. A simple martingale argument shows that τ_K can be any Markov stopping time; for example, the first regeneration time after some fixed number of iterations (Mykland, Tierney, and Yu 1992).

Before leaving this issue, we should explain a curiously attractive error. It seems natural to look at the estimates of probabilities Z_k/N_k obtained from single batches. These vary widely and seem to say something about the sampling variability, but they do not. Nothing is known about the distribution of Z_k/N_k; in particular, its expectation is not the probability of interest, because $E(Z_k/N_k) \neq E(Z_k)/E(N_k)$. The distribution of the tour lengths N_i will generally have a long tail; the few long tours contribute most of the information. This is an unavoidable consequence of stationarity and slow mixing of the cold chain. If each tour looks only at a small region of the state space, then the only way the stationary distribution can be correct is if tours that enter the cold chain in high probability regions are much longer than tours that enter in low probability regions. Any attempt to shorten the tail of the distribution of tour lengths must introduce bias.

Regeneration using an independence hot chain is not a necessary part of simulated tempering; it was not used by Marinari and Parisi (1992). But there is no way to know where it is safe to stop heating the distributions short of the "infinitely hot" independent sampling. When the sampler for the cold distribution alone would be very slowly mixing, it is the regeneration that provides all the mixing. There is no point to a simulated tempering sampler that does not make many excursions from end to end of the temperature range. So it is necessary to look at "tours" whether or not the sampler is regenerating. Despite this, if one knows either from theory or experience that a simulated tempering sampler without an independence hot chain mixes well, then regeneration should not be used, because, all other things being equal, the fewer distributions the better. We usually do not have such knowledge; it is safer to use regenerating samplers. Note that one need not have an independence hot chain to use regeneration; regeneration could be obtained by "splitting" the hot chain (Mykland et al. 1992), but we have not tried this.

3. THE WITCH'S HAT DISTRIBUTION

The "witch's hat" distribution in two dimensions is the distribution on the unit disc with a density shaped like a witch's hat, with a broad flat brim and a high conical peak. It was proposed by Matthews (1993) as a counterexample to the Gibbs sampler. In higher-dimensional analogs of the two-dimensional distribution, the mixing time of the Gibbs sampler increases exponentially with dimension, because all but one coordinate must be lined up with the peak before a Gibbs step can move from the brim to the peak, and this has exponentially small probability.

Here we use for illustration a simplified witch's hat distribution defined as follows. Let α and β be real numbers with $0 < \alpha \leq 1$ and $\beta \geq 0$. Define a distribution on the unit hypercube in d dimensions $[0, 1]^d$ as follows. The unnormalized density is equal to $1 + \beta$ on the small hypercube $[0, \alpha]^d$ equal to 1 elsewhere in $[0, 1]^d$. We still call the part of the distribution over the small hypercube the "peak" and the rest the "brim," although the density no longer looks much like a witch's hat. These distributions for various values of the parameters α and β make up the simplified witch's hat family. A hot distribution is the uniform distribution on the unit hypercube; $\alpha = 1$ or $\beta = 0$. For our main example, we chose $d = 30$ and a cold distribution with $\alpha = \frac{1}{3}$ and $\beta \approx 10^{14}$ chosen so the probability of the peak was exactly $\frac{1}{3}$.

A Gibbs sampler for this cold distribution has a very hard time. The peak is an atom, so the Gibbs sampler is regenerating. By the renewal theorem, the mean regeneration time is $1/P(\text{peak}) = 3$. The probability of leaving the peak in one scan of the Gibbs sampler is 6×10^{-13}. For the average time for tours of all lengths to be 3, the average length of tours of length greater than one must be 3.4×10^{12}. This characterizes the mixing of the Gibbs sampler. It will need 10^{12} scans to get close to mixing and 10 or 100 times that number to get any accuracy in the answers.

Some form of heating is necessary, but for the witch's hat, powering up is useless. Raising the cold distribution to a power still produces a distribution with two levels—the peak and the brim—in the same positions, so powering up is the same as decreasing β while leaving α fixed. This makes the peak no easier to hit and thus gives no improvement over ordinary Gibbs sampling. If the hot distribution has $\beta = 0$, then it is a regeneration point, so regeneration methods can be used to estimate variance. The overall acceptance rates will be high, but almost all tours will stay in the brims of the distributions. Over a very long run of the sampler there will eventually be a transition from the brim to the peak of some distribution, and then the sampler will stay in the peaks for 10^{12} iterations. Until such a long tour is seen, the regeneration estimates of variance will be completely erroneous.

Whereas the mixing time of the Gibbs sampler increases exponentially in d, the simulated tempering sampler needs a number of temperatures that is $O(d)$, and the mixing time is approximately quadratic in the number of temperatures (sec. 2.2), so the mixing time is approximately $O(d^2)$ if one counts iterations and $O(d^3)$ if one counts computing time, because the time to do one iteration is order d. This dependence is shown for the range $d = 30$ to $d = 300$ in Table 1.

For $d = 30$ we used the 22 temperatures shown in Table 2. The α's for intermediate temperatures were chosen to be equally spaced on the log scale, so that the area of each peak is the same fraction (.20816) of the peak for the next higher temperature. Thus there is a constant proportion of proposals in the peak in attempted jumps down in temperature. The β's were chosen so that the probability of the peak was equal to α. Because the hot distribution permits independent sampling, the sampler is regenerating. For this example we used

Table 1. Dependence of Computing Time on Dimension

d	m	time	n_{iter}	s	tours	tour len.
30	22	79.3	1,001,437	.0297	470	2,130.7
60	43	646.2	4,011,400	.0329	408	9,831.8
90	64	2,140.8	9,008,459	.0319	398	22,634.3
120	85	5,123.2	16,011,375	.0338	391	40,949.8
150	106	9,603.1	25,043,995	.0308	389	64,380.4
180	127	17,013.0	36,099,890	.0341	371	97,304.3
210	148	27,412.2	49,040,398	.0368	342	143,393.0
240	169	40,028.7	64,293,751	.0338	350	183,696.4
270	190	56,723.3	81,292,047	.0334	342	237,696.0
300	211	77,828.1	100,357,250	.0336	363	276,466.3

NOTE. d is dimension (number of variables), m is the number of distributions, time is running time in seconds done on a workstation about five times faster than those used for the rest of the computations, n_{iter} is the number of iterations (set proportional to d^2), s is the standard error of the estimator of the probability of the peak of the cold distribution, tours is the number of tours, and tour len. is the average length of a tour defined here to go from a regeneration until the next time the hot distribution is hit after hitting the cold distribution.

a pseudoprior that was exactly equal to the inverse normalizing constants $1/(1 + \beta\alpha^d)$.

The simulated tempering sampler was run to the first regeneration point after 1,000,000 iterations, which was iteration 1,000,110. This took 5 minutes and 42 seconds on a workstation that does about 1.5 million floating point operations per second. There were 42,556 tours of which all but 5,567 were of length 1 (regenerations on consecutive iterations). The distribution of the regeneration times was skewed (of course) but not extremely long tailed. The longest tour (11,556 iterations) made up only 1% of the total iterations. The largest 17 tours made up 10%, the largest 165 made up 50%, and the largest 773 made up 90%. The simulated tempering sampler gets one significant figure accuracy in about 10^6 scans. The exact results are given in Table 2.

Acceptance rates for jumps of the simulated tempering sampler are shown in Table 3. These acceptance rates are much larger than the recommendations in Section 2.2 at the cold end, but they cannot be made as small as 20% to 40%. Going down between temperatures 2 and 1, for example, the probability at stationarity of being on the brim before the jump is $1 - \alpha = .65$. When on the brim, the probability of a proposal on the brim is nearly 1, giving a contribution to the overall acceptance rate of 65% for jumps down in temperature at points on the brim of both distributions. The probability of being in the peak before the jump is .35, and the probability of a proposal in the peak is 20.8%. Most such proposals are accepted, giving a contribution to the overall acceptance rate of 7.3% for jumps down in temperature at points in the peak of both distributions. So although there is an overall acceptance rate of 72%, only 7% of that is involved in simulating the peak of the cold distribution.

Table 2. Results for the Simplified Witch's Hat Distribution ($d = 30$)

α	β	$\hat{\mu}$	Actual error	Estimated error
.333	1.03×10^{14}	.335	.001	.031
.351	2.32×10^{13}	.354	.003	.031
.370	5.24×10^{12}	.373	.003	.031
.390	1.19×10^{12}	.382	−.008	.030
.411	2.70×10^{11}	.403	−.008	.030
.433	6.14×10^{10}	.424	−.009	.029
.456	1.41×10^{10}	.441	−.016	.028
.481	3.23×10^{9}	.458	−.023	.027
.507	7.45×10^{8}	.486	−.021	.026
.534	1.73×10^{8}	.510	−.023	.024
.562	4.04×10^{7}	.541	−.021	.022
.593	9.52×10^{6}	.570	−.023	.021
.624	2.27×10^{6}	.607	−.018	.019
.658	5.46×10^{5}	.642	−.016	.017
.693	1.34×10^{5}	.676	−.017	.015
.731	3.33×10^{4}	.715	−.016	.013
.770	8.55×10^{3}	.759	−.011	.010
.811	2.28×10^{3}	.810	−.002	.007
.855	6.46×10^{2}	.855	0	.005
.901	1.99×10^{2}	.903	.002	.003
.949	6.98×10^{1}	.948	−.001	.001
1.000	0	1.000	0	0

NOTE: The cold distribution is the top row and the hot distribution the bottom; α and β are the parameters of the witch's hat distribution, μ is the probability of the peak, which is equal to α for the β values chosen here, $\hat{\mu}$ is the estimate of μ obtained by averaging over the samples. The "actual error" is the difference between $\hat{\mu}$ and $\mu = \alpha$. The "estimated error" is the standard error of $\hat{\mu}$ estimated using the ratio estimator.

4. LIKELIHOOD INFERENCE FOR THE STRAUSS PROCESS

The Strauss process (Strauss 1975) is the simplest non-Poisson Markov spatial point process. Here we deal with the

Table 3. Acceptance Rates for the Samples for the Simplified Witch's Hat Distribution

Temperature gap	Going up	Going down
1 to 2	.718	.720
2 to 3	.702	.707
3 to 4	.704	.690
4 to 5	.684	.676
5 to 6	.665	.659
6 to 7	.652	.637
7 to 8	.643	.627
8 to 9	.615	.615
9 to 10	.596	.594
10 to 11	.575	.570
11 to 12	.554	.546
12 to 13	.518	.519
13 to 14	.495	.496
14 to 15	.468	.467
15 to 16	.433	.431
16 to 17	.402	.400
17 to 18	.363	.363
18 to 19	.322	.322
19 to 20	.280	.286
20 to 21	.261	.263
21 to 22	.262	.260

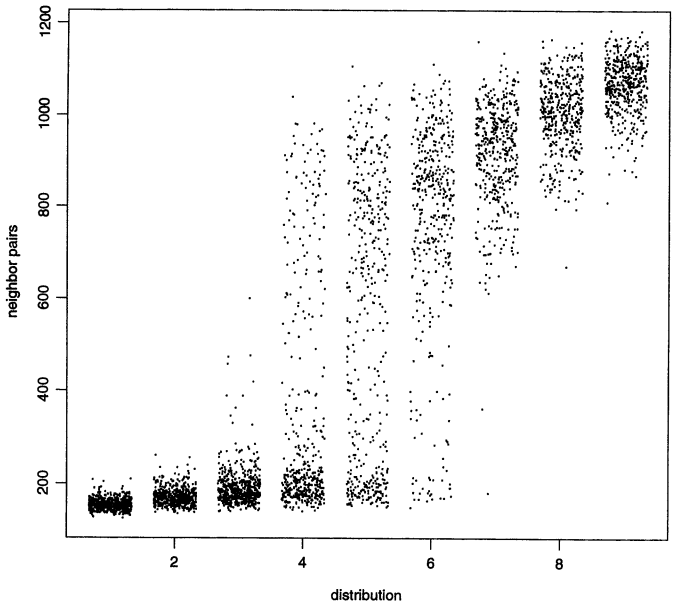

Figure 1. Scatterplot of the Canonical Statistics Versus Distribution Number for the Strauss Process. The x-coordinates are integer valued, but jittered. Every 100th iteration from a run of 405,677 iterations is plotted.

conditional Strauss process, which has realizations consisting of a fixed number of points in a bounded region. Let $t(x)$ denote the number of pairs of points (called *neighbor pairs*) separated by less than some fixed number r. A conditional Strauss process is any distribution in the exponential family with unnormalized densities $h_\theta(x) = e^{t(x)\theta}$ with respect to the "binomial process" under which the n points are uniformly distributed. Our example has 50 points in the unit torus and $r = .2$.

The first sampler for the conditional Strauss process was a Gibbs sampler (Ripley 1979). A Metropolis sampler described by Geyer and Møller (1994) is much more efficient and, unlike the Gibbs sampler, can be used for both the unconditional and conditional processes. Even the Metropolis algorithm is inefficient for a process with strong dependence (i.e., large positive θ). A simulated tempering sampler is better.

The special case $\theta = 0$ is the binomial process, which can be sampled independently and is a regeneration point. As θ increases so does the expected number of neighbor pairs, and for large θ all of the points are in one small clump and the value of $t(x)$ is very near its maximum $\binom{50}{2} = 1{,}225$ with very high probability. Preliminary runs showed that this occurs for $\theta > .16$, so we adjusted a sampler to have nine distributions, $\theta = 0, .0869, .1143, .1240, .1267, .1296, .1348, .1448, .16$, with approximately equal acceptance rates ranging between 65% and 77%. The results are shown in Figure 1.

We ran for 405,677 iterations, making 46,166 tours between regenerations, with 90 tours hitting the cold chain. The running time was 2 hours and 23 minutes on a workstation that does about 1.5 million floating point operations per second. This one sample describes this conditional Strauss process for all values of θ between 0 and .16. In particular the mapping between the canonical parameter θ and the mean value parameter $\tau(\theta) = E_\theta t(X)$ can be determined by importance reweighting the sample. Let X_k, I_k denote the samples, which have unnormalized stationary density $h_{\theta_i}(x)\pi(i)$, and let

$$w_\theta(x, i) = \frac{h_\theta(x)}{h_{\theta_i}(x)\pi(i)}.$$

Then

$$\tau_n(\theta) = \frac{\sum_{k=1}^n t(X_k) w_\theta(X_k, I_k)}{\sum_{k=1}^n w_\theta(X_k, I_k)} \to \tau(\theta), \qquad n \to \infty \quad (4)$$

for each θ, and $\tau_n(\theta)$ is the natural Monte Carlo approximation of $\tau(\theta)$. This curve is shown in Figure 2. Maximum likelihood estimation is now a simple matter of finding the θ such that $\tau_n(\theta)$ equals the observed $t(x)$. Monte Carlo likelihood to theory applies here just as to any other Markov chain sampler. The novelty is in the faster mixing, which allows Figure 2 to be computed easily (cf. Strauss 1986).

5. ANCESTRAL INFERENCE IN THE HUTTERITES

5.1 The Genetic Model

We consider the inheritance at single diallelic genetic locus. This means that each individual has two genes, and that there are two types of genes (*alleles*), denoted by A and a. Hence each individual has one of three possible genotypes: AA, Aa, or aa. We consider a lethal recessive disease; that is, the AA and Aa genotypes produce individuals with *normal* characteristics, but all aa individuals die before the age of reproduction. Conversely, all individuals who have survived to adulthood (and, in particular, any parent) must be either genotype AA or Aa (called *noncarrier* or *carrier*). The problem of interest is to compute the probability distribution of carrier status over the pedigree given the observed data.

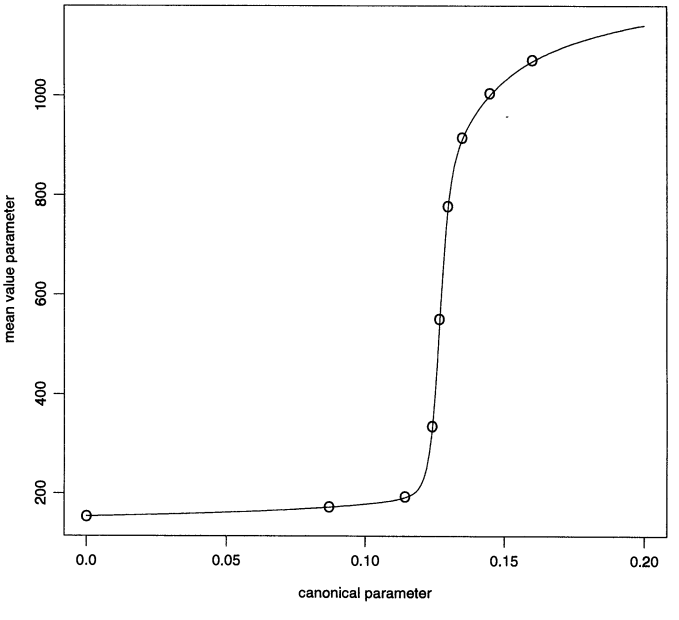

Figure 2. Plot of the Mean Value Parameter $\tau(\theta)$ Versus the Canonical Parameter θ for the Strauss Process. The dots are the empirical averages for the nine distributions sampled. The line is $\tau_n(\theta)$ given by Equation (4).

Mendel's laws specify the probability of an individual's genotype given the genotypes of the parents. If neither parent is a carrier, then the child must be AA. If one parent is a carrier, then the child has probability .5 of being Aa and .5 of being AA. If both parents are carriers, then the probabilities are .25, .5, and .25 of the child being AA, Aa, or aa. Individuals whose parents are unknown (founders) are assumed to have genes that are a random draw from the population gene pool. Their genotype probabilities are given by

$$\Pr(\mathrm{AA}) = (1-p)^2,$$
$$\Pr(\mathrm{Aa}) = 2p(1-p), \qquad \Pr(\mathrm{aa}) = p^2, \tag{5}$$

where p is the population frequency of the disease gene (assumed known). This specifies the probabilities in the model.

Tracing the ancestry of rare recessive diseases in genetic isolates has been considered often (see, for example, Castilla and Adams 1990, Hussels and Morton 1972, Sorsby 1963, and Thompson and Morgan 1989). But except where an exact probability can be computed (Thompson 1978), the methods used are of doubtful value. On a large complex pedigree, exact computation of posterior probabilities is infeasible. Although the Gibbs sampler (Geman and Geman 1984) has been used successfully to estimate probabilities on small pedigrees, on the pedigree of our example the Gibbs sampler does not mix, even in very long runs (M. Emond, unpublished results). For large pedigrees, methods like the Gibbs sampler that update one variable at a time can take eons to get a representative sample of genotypic configurations.

5.2 The Genealogy and Cystic Fibrosis

We illustrate our methods with a problem that has stretched them to their limits: the ancestry of cystic fibrosis (CF) genes in the Hutterite population of North America. The structure of this large Caucasian genetic isolate has been described by Hostetler (1974), and the CF data has been described by Fujiwara et al. (1988). The current population of more than 30,000 traces its entire ancestry to about 85 founders mostly living in the eighteenth century. About 450 immigrants came to North America in the late nineteenth century, and the population expanded very rapidly thereafter. CF is a recessive and (until recently) lethal genetic disease. The frequency of CF genes in Caucasian populations is typically about .025; in large Caucasian populations, about 1 in 1,600 births is affected by CF, and about 1 in 20 individuals is a carrier. This gene frequency seems plausible for the founders of the Hutterite population, although, due to genetic drift and founder effects, the frequency in the current population may be higher.

In the data set we consider, 27 couples are known to be carriers because they are parents of diagnosed CF cases (K. Morgan, personal communication). These 54 known carrier parents, together with all their direct ancestors tracing back to the orginal founders, number 771. These founders, the majority of whom lived before 1750, number 77. This is the *core pedigree.* The data base of Hutterite individuals born up to 1981 (T. M. Fujiwara and K. Morgan, unpublished data) contains 24,875 individuals. Analysis of this entire population pedigree is feasible but would require huge amonts of computing time. But an analysis of CF ancestry based only on the core pedigree would be biased. The ancestors of current cases had many other descendants who lived to adulthood and thus cannot have been affected by CF.

First, we restricted attention to the offspring of members of the core pedigree. There are 1,242 such offspring who themselves had offspring and so can be assumed to be unaffected. Adding them and their 11 additional parents not in the core pedigree makes a 2,024-member pedigree, which is the subject of our main analysis. We later analyzed a larger pedigree of 5,277 individuals, adding to the core pedigree all the children and grandchildren of the core pedigree who themselves had offspring (and thus can be assumed unaffected).

In computing probabilities on pedigrees, it is often convenient to preprocess information from the periphery of the structure (Thompson 1978), and such contributions to the overall result can be incorporated into MCMC sampling on the remainder (Thompson 1991). Here we replace children with no offspring by *pair potentials* on their parents. Let x be the genotype of such a child and let x_m and x_f be the genotype of the child's parents. Then the contribution to the probability distribution for this child is the pair potential

$$\phi(x_m, x_f) = \sum_x \Pr(\text{data on the child} \mid x)\Pr(x \mid x_m, x_f).$$

The marginal probability distribution for the remaining individuals is simply the distribution for the rest times the product of the pair potentials. For the Hutterites, this greatly decreases the amount of work the sampler must do. In our 2,024-member pedigree, 1,209 individuals have no offspring in this pedigree and can be replaced by pair potentials on their parents. This leaves only 815 individuals to be sampled. The sampler not only takes less than half the time to make one scan but is also less sticky, because the potentials provide part of the distribution exactly. In the 5,277-member pedigree, 3,167 individuals were replaced by pair potentials, leaving 2,110 individuals actually sampled.

5.3 Hot Distributions and Hot Priors

The regeneration method needs a "hot" distribution, h_m, for which independent sampling is feasible. For our pedigree analysis problems, we used two different distributions for independent sampling: *gene-drop* and *all-carriers.* Gene-drop is the distribution of the genotypes when the data are ignored. It is easily simulated by drawing the founders' genotypes independently from Equation (5), then going down the pedigree simulating offspring genotypes conditionally on their parents'. All-carriers is the distribution that gives probability 1 to the genotypic configuration in which every individual is a carrier, Aa. (The cases who are known to have genotype aa are not in the 2,024-member or 5,277-member pedigrees.) This distribution is even easier to simulate; every realization is the same.

There is no reason not to change other aspects of the model as well. We also experimented with individual-specific "hot priors," changing the prior distribution for certain founders so that the gene-drop would make them carriers more fre-

quently. Adjusting the hot priors so that the founders have approximately the same carrier frequencies in both the hot and cold distributions makes the sampler more efficient but requires some iteration. Note that the hot priors do not alter the cold distribution; the sampler mixes faster with good hot priors, but it produces valid results regardless.

Either of these two hot distributions can be thought of as resulting from altering the penetrances (Pr(data | genotype)). The gene-drop distribution results from a uniform penetrance over all data values for each genotype, and the all-carriers distribution results from zero penetrance of the AA and aa genotypes. For "warm" distributions intermediate between hot and cold, we used penetrances that were convex combinations of the hot and cold (true) penetrances, λ of the hot penetrances, and $1 - \lambda$ of the cold penetrances, where $0 \le \lambda \le 1$. When hot priors were used, the warm distributions had similar convex combinations of the hot and cold priors.

5.4 Results

The results of our analysis of the 2,024-member pedigree are shown in Figure 3 and the first two columns of Table 4. Figure 3 gives a histogram of all the carrier probabilities. The prior (unconditional) probability of being a carrier is .049. Of the 77 founders of the core pedigree, one is a known carrier. Of the other 76 founders, 45 are more than two standard errors (of the Monte Carlo) from the prior mean; 12 below and 33 above. A few founders are far above the unconditional probability; the 13 with the highest carrier probabilities (as estimated by the Monte Carlo) are shown in Table 4. Their probabilities of being carriers range from almost two to more than four times the prior probability. Note that the couples C-D, E-F, G-H, and I-J, who must have exactly the same true carrier probabilities, have Monte Carlo estimates that agree to within the estimated Monte Carlo error. The conditional expectation of the number of CF genes in these 76 founders is 5.58 (standard error .05); the unconditional expectation is 3.705.

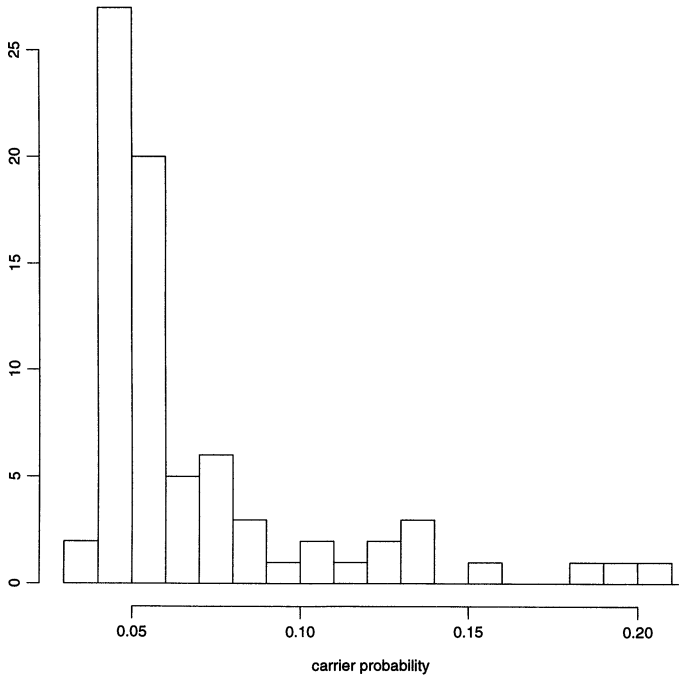

Figure 3. Histogram of the Estimated Carrier Frequencies for the 76 Founders of the Core Pedigree Who Were not Known Carriers.

Table 4. Hutterite Carrier Frequencies

	2,024 members		5,277 members	
	Mean	*S.E.*	*Mean*	*S.E*
A	.204	.005	.318	.024
B	.195	.015	.294	.031
C	.183	.014	.088	.021
D	.159	.011	.089	.023
E	.140	.013	.140	.019
F	.134	.013	.109	.015
G	.133	.014	.076	.011
H	.127	.012	.071	.009
I	.121	.008	.164	.015
J	.116	.008	.163	.016
K	.109	.011	.073	.016
L	.104	.009	.063	.011
M	.094	.014	.060	.007

NOTE: 2,024 members refers to the pedigree containing ancestors of affected individuals and their first-generation offspring who themselves had offspring and are thus known to not have CF. 5,277 members refers to the pedigree containing ancestors of affected individuals and their first- and second-generation descendents who themselves had offspring. Mean is the estimated posterior probability of being a carrier, and S.E. is the Monte Carlo standard error of the estimate. The first column gives arbitrary labels for the individuals. The pairs C-D, E-F, G-H, and I-J are married couples with no other spouses.

These estimates were based on a run of 11,555,470 iterations (each iteration being one Gibbs scan of the 815 individuals being sampled plus an attempt to jump from one distribution to another), during which there were 355 tours that spent any time sampling the distribution of interest. The total running time was 20 days, 3 hours on a workstation that does about 2 million floating point operations per second.

The standard errors are based on the sampling variability of these 355 tours. The distribution of tour lengths is shown in Figure 4. The tours range in length from 1 to 8,830 and approximately follow Zipf's law: 35 tours account for half of the total length, another 38 account for half of the remaining half, another 37 for half of the remaining quarter, another 34 for half of the remaining eighth, another 29 for half of the remaining sixteenth, and so forth.

The estimation for a single individual is illustrated in Figure 5, which shows the results of the Monte Carlo for individual "B" in Table 4, who was chosen because he or she had higher carrier probability and also large Monte Carlo error (being at the top of the pedigree). The slope of the line in the figure is the sum of all the y values of the points divided by the sum of all the x values. So the points cluster around the line, but not in any very obvious sense; the long tours provide most of the information.

The operating characteristics of the sampler are shown in Figures 6 and 7. Figure 6 shows the occupation numbers as a function of the parameter λ, indexing the distributions. Figure 7 shows the acceptance rates, which were adjusted to a desired acceptance rate of 40%. In neither case was the adjustment perfect, the deviations from uniformity being larger than the sampling variability, but the misadjustment does not seriously degrade performance.

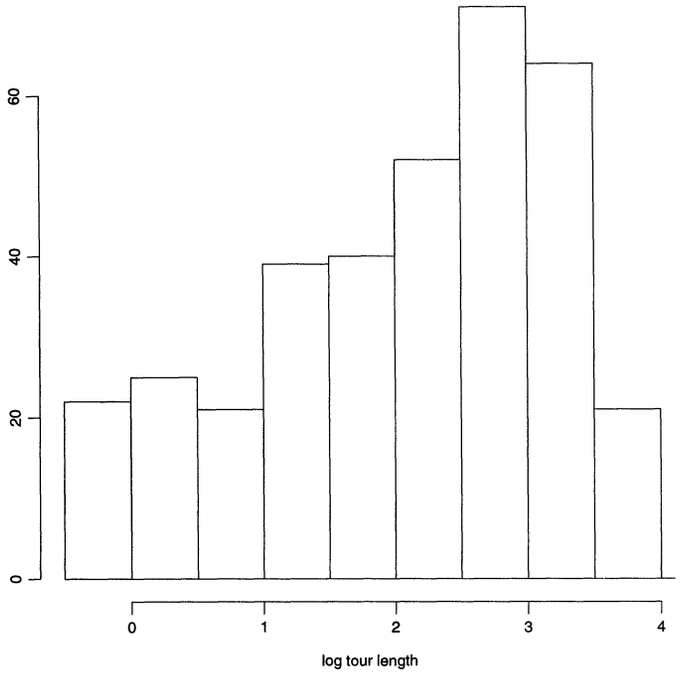

Figure 4. Histogram of Log (Base 10) of Tour Lengths in the Monte Carlo for the Hutterite Pedigree.

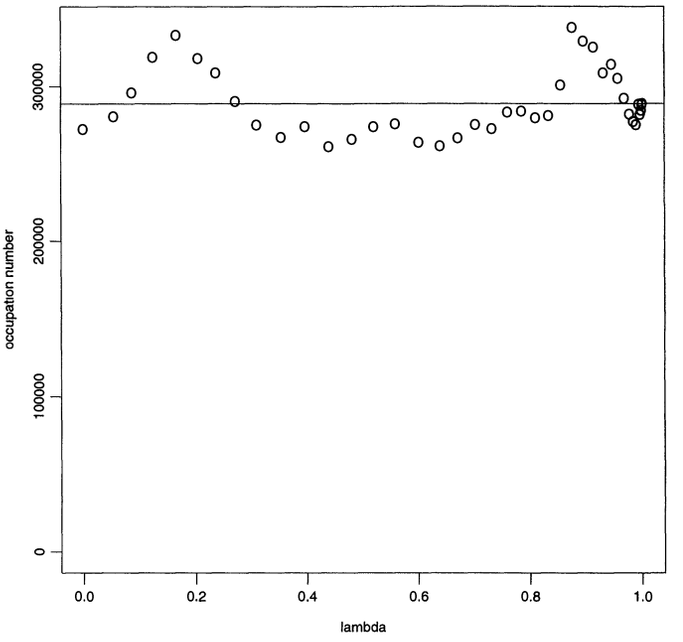

Figure 6. Occupation Number for the Hutterite Pedigree Sampler.

Using the information from this run, the pseudopriors and λ's were adjusted to get uniform occupation numbers and acceptance rates of 30%. This sampler had 32 distributions. A run of 2,255,775 iterations showed that the adjustment was fairly successful. The occupation numbers were perhaps uniform to within sampling error, and the acceptance rates were all 30% or 31%, except for three of 29, 33, and 35. This sampler appeared to run about 8% faster than the other, but that may have been only sampling variation. Then another sampler with 26 distributions was adjusted to have acceptance rates of about 20%. A run of 2,008,438 iterations showed almost uniform occupation numbers and acceptance rates all between 19% and 21%, except for three of 18, 22, and 26. This sampler appeared to run about 5% faster still. Although the sampling error in the speeds of the

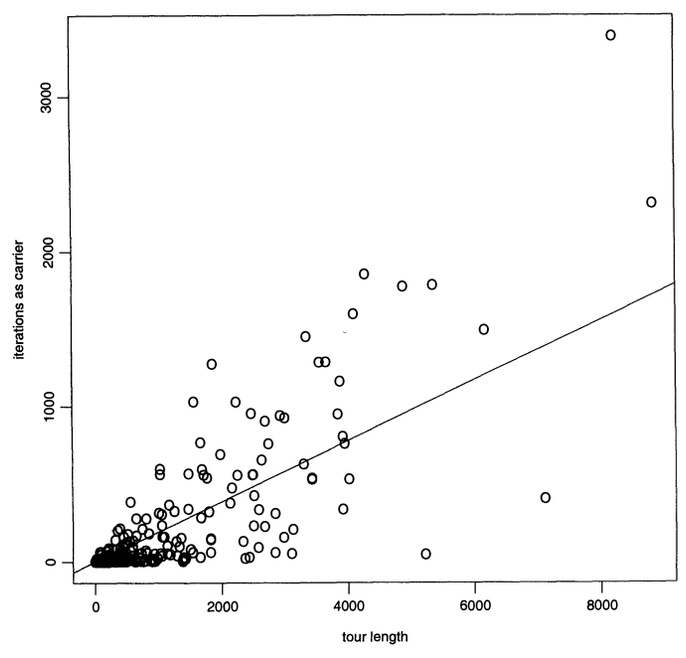

Figure 5. Scatterplot of Number of Iterations During a Tour That Individual "B" was a Carrier Against Tour Length. The line goes through the origin and has slope equal to the estimated carrier frequency for individual "B."

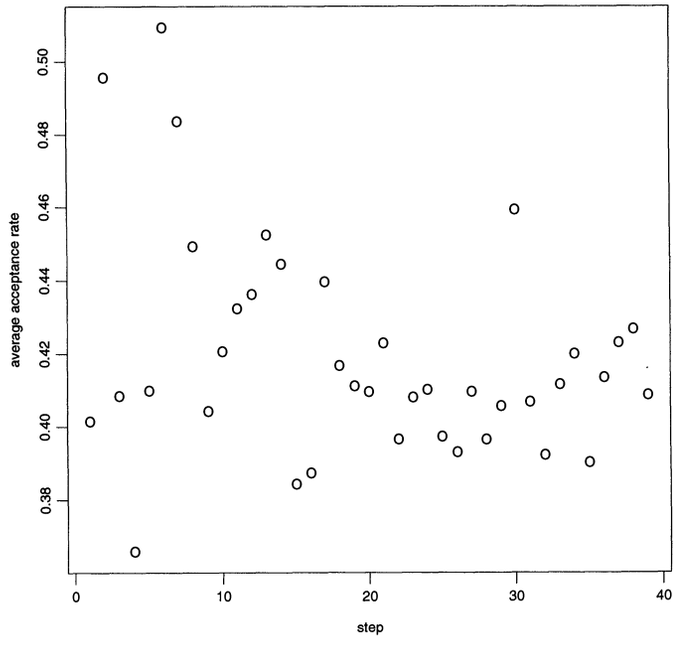

Figure 7. Plot of the Average Acceptance Rate of Jumps Between Distributions. The average is the average of the acceptance rate going up and the acceptance rate going down. At the ends, the acceptance rates were adjusted to account for the uneven proposal probabilities. The "steps" are numbered from 1 to 39 going from cold to hot.

latter two samplers is fairly large, adjusting the acceptance rates to between 20% and 40% seems reasonable.

Results on the 5,277-member pedigree are shown in the second two columns of Table 4. This sampler had 42 distributions and ran for 12,314,658 iterations, producing 37 tours that hit the cold distribution. The tours for this sampler are about 10 times the length of tours for the 2,024-member pedigree. Because of the smaller number of tours, this sampler is less accurate than the one for the 2,024-member pedigree, but it is accurate enough to show that the two pedigrees do have different probability distributions. Individuals A and B are now much more likely than the others to have been carriers, half again as likely as given the information in the 2,024-member pedigree. Presumably the answer for the full pedigree has a still higher probability of A and B being carriers.

6. DISCUSSION

For the purposes of discussion, let us divide problems into "hard" ones that need simulated tempering and "easy" ones for which the Gibbs sampler or variable-at-a-time Metropolis algorithms work. The main value of simulated tempering is that it provides a method of attack for these "hard" problems. The method is not guaranteed, because if one chooses a bad form of "heating," simulated tempering can fail (Sec. 3). But no other MCMC method has guaranteed convergence either, and simulated tempering seems to provide the best chance of obtaining a converging sampler in hard problems.

In easy problems the function of simulated tempering is to remove doubts about convergence of the Gibbs sampler and other simple methods. If simulated tempering produces the same answer as the simpler methods, then both presumably are right. There has been much controversy in the literature over the convergence even of very simple examples (Gelman and Rubin 1992; Geyer 1992). In such cases a solution is to run simulated tempering, which seems to deliver the benefits that were promised for multistart methods (Gelman and Rubin 1992). Figure 5 shows why multistart methods will not work in hard problems. A multistart method would produce some average over the dots in the figure that would depend on the starting distribution and hence be incorrect unless the starting distribution was very near the stationary distribution.

Have we found effective hot distributions for the Hutterite CF problem? The sampler found "modes" in which each founder was a carrier, so it could have missed a mode only if the mode were characterized by some more complex function of the paths of descent of the CF genes. We used two different hot distributions. The results for the gene-drop hot distribution have not been shown, but agreed with those discussed to within the estimated Monte Carlo error. Because no other method that we know of mixes well enough to check our results on this large pedigree, we cannot guarantee the results are correct, but available evidence suggests that they are.

[Received July 1993. Revised November 1994.]

REFERENCES

Berg, B., and Neuhaus, T. (1991), "Multicanonical Algorithms for First Order Phase Transitions," *Physics Letters B,* 267, 249–253.

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 55, 25–37.

Cannings, C., Thompson, E. A., and Skolnick, M. H. (1978), "Probability Functions on Complex Pedigrees," *Advances in Applied Probability,* 10, 26–61.

Castilla, E. E., and Adams, J. (1990), "Migration and Genetic Structure in an Isolated Population in Argentina: Aicuna," in *Convergent Issues in Genetics and Demography,* eds. J. Adams, A. Hermalin, D. Lam, and P. Smouse, New York: Oxford University Press.

Feller, W. (1968), *An Introduction to Probability Theory and Its Applications,* Vol. 1 (3rd ed., rev.), New York: John Wiley.

Frantz, D. D., Freeman, D. L., and Doll, J. D. (1990), "Reducing Quasi-Ergodic Behavior in Monte Carlo Simulations by J-Walking: Applications to Atomic Clusters," *Journal of Chemical Physics,* 93, 2769–2784.

Fujiwara, T. M., Morgan, K., Schwarz, R. H., Doherty, R. A., Miller, S. H., Klinger, K., Stanislovitis, P., Stuart, N., and Watkins, P. C. (1988), "Genealogical Analysis of Cystic Fibrosis and Chromosome 7q RFLP Haplotypes in the Hutterite Brethren," *American Journal of Human Genetics,* 44, 327–337.

Gelman, A., and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science,* 7, 457–511.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721–741.

Geyer, C. J. (1991a), "Markov Chain Monte Carlo Maximum Likelihood," in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* pp. 156–163.

——— (1991b), "Reweighting Monte Carlo Mixtures," Technical Report No. 568, University of Minnesota, School of Statistics.

——— (1992), "Practical Markov Chain Monte Carlo" (with discussion), *Statistical Science,* 7, 437–511.

Geyer, C. J., and Møller, J. (1994), "Simulation and Likelihood Inference for Spatial Point Processes," *Scandinavian Journal of Statistics,* 21, 359–373.

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 54, 657–699.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika,* 57, 97–109.

Hostetler, J. A. (1974), *Hutterite Society,* Baltimore: Johns Hopkins University Press.

Hussels, I. E., and Morton, N. E. (1972), "Pingelap and Mokil Atolls: Achromatopsia," *American Journal of Human Genetics,* 24, 304–309.

Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science,* 220, 671–680.

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

Matthews, P. (1993), "A Slowly Mixing Markov Chain With Implications for Gibbs Sampling," *Statistics and Probability Letters,* 17, 231–236.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics,* 21, 1087–1092.

Mykland, P., Tierney, L., and Yu, B. (1995), "Regeneration in Markov Chain Samplers," *Journal of the American Statistical Association,* 90, 233–246.

Nummelin, E. (1984), *General Irreducible Markov Chains and Non-Negative Operators,* Cambridge, U.K.: Cambridge University Press.

Ripley, B. D. (1979), "Simulating Spatial Patterns: Dependent Samples From a Multivariate Density," *Applied Statistics,* 28, 109–112.

Ripley, B. D. (1987), *Stochastic Simulation,* New York: John Wiley.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 55, 3–23.

Sorsby, A. (1963), "Retinitis Pigmentosa in the Tristan da Cunha Islanders," *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 57, 15–18.

Strauss, D. J. (1975), "A Model for Clustering," *Biometrika,* 62, 467–475.

Strauss, D. (1986), "A General Class of Models for Interaction," *SIAM Review,* 28, 513–527.

Swendsen, R. H., and Wang, J. S. (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physical Review Letters,* 58, 86–88.

Thompson, E. A. (1978), "Ancestral Inference II: The Founders of Tristan da Cunha," *Annals of Human Genetics,* 42, 239–253.

——— (1991), "Probabilities on Complex Pedigrees; the Gibbs Sampler Approach," in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* pp. 321–328.

Thompson, E. A., and Morgan, K. (1989), "Recursive Descent Probabilities for Rare Recessive Lethals," *Annals of Human Genetics,* 53, 357–374.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *Annals of Statistics,* 22, 1701–1762.

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.

Wang, J. S., and Swendsen, R. H. (1990), "Cluster Monte Carlo Algorithms," *Physica A,* 167, 565–579.

Wasan, M. T. (1969), *Stochastic Approximation,* Cambridge, U.K.: Cambridge University Press.