Political Analysis (2004) 12:425–443
doi:10.1093/pan/mph027

Dynamic Tempered Transitions for Exploring Multimodal Posterior Distributions

Jeff Gill
Department of Political Science, University of California, Davis,
One Shields Avenue, Davis, CA 95616
e-mail: jgill@ucdavis.edu

George Casella
Department of Statistics, University of Florida, Griffin-Floyd Hall,
P.O. Box 118545, Gainesville, FL 32611
e-mail: casella@stat.ufl.edu

Multimodal, high-dimension posterior distributions are well known to cause mixing problems for standard Markov chain Monte Carlo (MCMC) procedures; unfortunately such functional forms readily occur in empirical political science. This is a particularly important problem in applied Bayesian work because inferences are made from finite intervals of the Markov chain path. To address this issue, we develop and apply a new MCMC algorithm based on tempered transitions of simulated annealing, adding a dynamic element that allows the chain to self-tune its annealing schedule in response to current posterior features. This important feature prevents the Markov chain from getting trapped in minor modal areas for long periods of time. The algorithm is applied to a probabilistic spatial model of voting in which the objective function of interest is the candidate's expected return. We first show that such models can lead to complex target forms and then demonstrate that the dynamic algorithm easily handles even large problems of this kind.

1 Introduction

1.1 *Background Purpose*

The advent of estimation based on Markov chain Monte Carlo (MCMC) revolutionized Bayesian inference beginning with the review essay of Gelfand and Smith (1990). Simulation essentially replaced analytical derivation of marginal posterior forms from joint expressions, allowing Bayesian researchers to develop more realistic and more complex model specifications. As a result, Bayesian posterior distributions in applied work have become more complex and high dimensional and are often found to have multiple modes.

It is well known that highly multimodal target distributions are problematic for basic MCMC algorithms. Such multimodal distributions provide a plethora of local maxima to attract and trap Markov chains for extended periods of time that can exceed reasonable

Authors' note: Our thanks to Alan Agresti, Jim Booth, Patrick Brandt, James Fowler, Jim Hobert, Christian Robert, and the National Science Foundation (DMS-99-71586). Software, written in R, to implement the method described in this paper is available at the *Political Analysis* Web site.

chain lengths and thus prevent a full exploration of the posterior distribution of interest. Furthermore, this *mixing problem* is greatly exacerbated with higher dimensions and with dependencies between variables. The problem is not unique to Bayesian stochastic simulation, as multiple modes also present difficulties for standard maximum likelihood numerical algorithms: quasi-Newton method BFGS, standard and modified Newton-Raphson, steepest descent, and so forth (see the recent extended discussion in Altman et al. 2004). Therefore an algorithmic solution that allows for more complete exploration of such complex forms has the potential to improve general statistical computing.

It is important to remember that mixing is a distinct problem from convergence, which has consumed much of the recent MCMC literature (although time to convergence is sometimes referred to as *mixing time*). A Markov chain has converged to its limiting distribution (the posterior of interest for properly set up MCMC applications) when it generates only legitimate values from this distribution in proportion to the actual target density. Conversely, mixing refers to the rate at which the Markov chain moves about the parameter space, either before or after convergence. Thus even a Markov chain that has converged to its limiting distribution may be of little or no value if it has not sufficiently mixed through the distribution because collected chain values will be incomplete and empirical summaries will be biased. Since Markov chains are run asymptotically only in texts, it is important to use a sampling strategy that leads to mixing in a time that is reasonable for the research project.

Critically, a fast-mixing Markov chain will reach convergence sooner than a corresponding version with slow mixing properties. Thus it is the researcher's responsibility to assess the thoroughness of mixing: the degree to which the full state space is explored. Formal tests for mixing can be quite involved, so most experienced practitioners compare the range of chain visits in each dimension to the known distributional support for that dimension.

In this article we first describe a lineage of MCMC estimation and description procedures starting with the foundational Metropolis-Hastings algorithm. A great deal of MCMC work is based on Metropolis-Hastings because it is more flexible than its primary competitor, Gibbs sampling. We then show how the simulated annealing variant works and why it is useful in cases with difficult posterior forms. Next we introduce a set of solutions based on simulated annealing: Metropolis coupled Markov chain Monte Carlo, simulated tempering, tempered transitions, and dynamic tempered transitions. The latter is the new approach that we developed out of dissatisfaction with the mixing properties of the more basic algorithms in the presence of multiple modes.

The reward for developing this more complex derivative of the basic Metropolis-Hastings algorithm is that it can provide full exploration of highly complex and multimodal posterior distributions within very reasonable chain lengths. Rasmussen (2003) calls these types of problems "expensive" because simple MCMC schemes provide impractically long periods for full posterior exploration. We know that such posterior forms arise in many disciplines, including political science, and have stymied much research.

The dynamic sampler is contrasted with others using an objective function that presents a particularly difficult mixing challenge, followed by an application to probabilistic voting theory. Complexity, obfuscation, vagueness, and uncertainty are permanent features of U.S. electoral politics. Both voters and candidates often have motivations to make their issue positions deliberately vague. Probabilistic spatial voting models address these and other concerns by explicitly including an error term that can accommodate limited probabilistic knowledge in vote calculations. For example, if $\mathbf{U}_{ij}$ is the utility of candidate j to voter i, then a deterministic model posits that this utility is a (possibly complex) negative function of the issue distance between the voter's preferred position and that of the candidate: $\mathbf{U}_{ij} = -f(\mathbf{D}_{ij})$. A probabilistic model includes an error term that can alter this basic calculus:

$\mathbf{U}_{ij} = \mathbf{E}_j - f(\mathbf{D}_{ij})$. The "cost" that one often pays for specifying more complicated forms with such specifications (i.e., including explicit error terms, hierarchical relationships, and dependent covariates) is an objective function with interesting and perhaps difficult features. The major purpose in our using this example is to show researchers that the new MCMC estimation tool possesses robust general properties in such difficult settings. Our code for dynamic tempered transitions is made freely available in the R statistical environment, so readers can readily apply the described algorithm to their own problems.

As the Bayesian approach gains in popularity in empirical political science, improved MCMC estimation procedures tailored to the types of problems we face in particular become more important. For instance, the application given here demonstrates that simple voting models for realistic numbers of participants, issues, and candidates can readily lead to complex aggregate preferences expressed in multimodal objective functions. The MCMC scheme developed here addresses these types of demanding functional forms and a great many more, since it is not deterred by complex shapes and high dimensions.

2 A Set of Developed Procedures

For generalized linear models, the maximum likelihood estimate for the unknown parameter vector typically does not have a closed-form analytical solution like that provided by ordinary least squares in the linear model context. So numerical mode-finding techniques implemented in software are required where the goal is to find the root of the first derivative function and evaluate the likelihood properties at this point. The primary iterative root-finding procedure implemented in general use and particularistic packages is *iterative weighted least squares* (IWLS), introduced by Nelder and Wedderburn (1972). This algorithm extends the standard Newton-Raphson strategy by iteratively reweighting a general least squares algorithm, improving the estimate on each cycle using the mean function. In almost all applications this procedure works extremely well and users are unconcerned about the details (occasionally things can go awry, requiring more complex procedures; see Gill and King [forthcoming]).

Such a procedure is unsatisfying for Bayesian inference because the goal there is to fully describe a multidimensional posterior form. Thus Bayesian procedures need to make use of numerical integration rather than numerical differentiation in order to give marginal descriptions and subsequent summaries of interest. Obvious and traditional methods of numerical integration include Newton-Cotes, Riemann approximations, Laplace approximations, and Gaussian quadrature. More modern variants include rejection sampling, saddlepoint approximation, and importance sampling. Technical details of these and other variants are given in the early chapters of Robert and Casella (1999).

Unfortunately such well-known numerical integration procedures often fail with reasonably complicated posterior forms because they require some knowledge about marginal characteristics in order to be set up. This problem shackled Bayesians for decades before the advent of Markov chain Monte Carlo as a means of generating empirical values from the marginals of interest (see the introduction to this special issue). In this article we will focus on one of the basic variants of MCMC procedures, the Metropolis-Hastings algorithm, because of its inherent customizability.

2.1 *Metropolis-Hastings*

First, define terms for a standard Metropolis-Hastings algorithm, the *random walk Metropolis-Hastings process.* The K-dimensional position of Markov chain j is the K-length vector $\mathbf{c}_j = [C_{j1}, C_{j2}, \ldots, C_{jK}]$, and we label the target (posterior) density as $\tau_j(\ )$ to

distinguish it from proposal densities. Multiple Markov chains are indexed here by j since it is common and justified to start multiple chains from differing origins as a means of robusting with regard to starting points and false attractions. Now at iteration t chain j takes on the value $\mathbf{c}_j^{(t)}$, and movement is governed by the Metropolis-Hastings steps:

1. Generate $\mathbf{c}_j' \sim m(\mathbf{c}_j)$ from a proposal distribution according to

$$m(\mathbf{c}_j) = \mathbf{c}_j^{(t)} + \xi, \quad \xi \sim \mathcal{N}(0, \sigma^2).$$

2. Calculate

$$\rho_t = \frac{\tau_j(\mathbf{c}_j')m(\mathbf{c}_j^{(t)})}{\tau_j(\mathbf{c}_j^{(t)})m(\mathbf{c}_j')}.$$

3. $\mathbf{c}_j^{(t+1)} = \begin{cases} \mathbf{c}_j' & \text{with probability } \min\{\rho_t, 1\} \\ \mathbf{c}_j^{(t)} & \text{otherwise} \end{cases}$

4. Cycle through $\mathbf{c}_{j=1}^{(t)}, \ldots, \mathbf{c}_{j=N}^{(t)}$ drawing a sample of size M for each chain.

Thus we get a combined sample of values of size $N \times M$, allowing us to examine not only modes but also the range of values reflecting the distribution of interest. In most applications this MCMC process works very well, converging to the distribution of interest, $\tau_j(\ )$, and mixing throughout.

2.2 *Simulated Annealing*

A standard alternative when Metropolis-Hastings mixing and convergence does not readily occur is to use *simulated annealing* (Kirkpatrick et al. 1983). Simulated annealing "melts down" the peaks of the distribution in order to facilitate easier movement of the Markov chain and therefore more thorough travel through the parameter space. This is done simply by raising the distribution function to some power and then systematically returning the exponent to one according to some prearranged cooling schedule. More specifically, at iteration t' we have the value $\mathbf{c}_j^{(t')}$ and perform the steps:

1. Generate $\mathbf{c}_j' \sim \left[\mathbf{c}_j^{(t')} + \mathcal{N}(0, \sigma^2)\right]$.
2. Calculate

$$\rho_{t'} = \exp\left(\frac{\tau_j(\mathbf{c}_j') - \tau_j(\mathbf{c}_j^{(t')})}{\beta_{t'}}\right).$$

3. $\mathbf{c}_j^{(t'+1)} = \begin{cases} \mathbf{c}_j' & \text{with probability } \min\{\rho_{t'}, 1\} \\ \mathbf{c}_j^{(t')} & \text{otherwise} \end{cases}$

The cooling sequence that generates values of $\beta_{t'}$ will typically be of the form

$$\beta_{t'} = \text{constant}/\log(t' + 1),$$

where the constant is chosen to insure adequate mixing at the start of the iterations. However, in practice researchers often "shortcut" this type of cooling schedule and use a faster scheme (geometric, exponential, linear, etc.).

Heating the kernel flattens out its probability structure toward a uniform distribution, and if there are many modes, they will melt into the surface and will therefore no longer be false attractions. As the jumping distribution generates candidate positions, very few of these will be rejected and the Markov chain will rarely stay in place. This is good; it means that the chain can freely explore the sample space without impediments. It is also bad in that there is obviously much less of a tendency to remain in the (previous) high-density areas. This is where care must be taken in determining the cooling process: slow cooling ensures greater coverage of the sample space, but faster cooling back to zero in order gives reasonable simulation times.

2.3 *Metropolis-Coupling*

With high-dimensional and multimodal objective functions (posteriors) of interest, the candidate distribution and the temperature schedule in simulated annealing must be chosen with great care to allow adequate exploration of the space. Unfortunately it is possible to stipulate wildly inappropriate choices of both, thus preventing convergence or mixing with simulated annealing. One early solution to this problem is *Metropolis-coupled Markov chain Monte Carlo* (MCMCMC) (Geyer 1991). This algorithm is characterized by the steps:

1. Run N parallel chains at different heat levels from m^1 to m^{1/β_N}, where the temperature values have the characteristic $\beta_1 = 1 < \beta_2 < \cdots < \beta_N$.
2. Thus N transition kernels are defined, MC_1, MC_2, $\ldots$, MC_N.
3. At time t select two chains, i and j, and attempt to swap states:

$$\mathbf{c}_i^{(t)} \Leftarrow \mathbf{c}_j^{(t)}, \qquad \mathbf{c}_j^{(t)} \Leftarrow \mathbf{c}_i^{(t)},$$

4. with a Metropolis decision probability:

$$\min\left\{1, \frac{m_i(\mathbf{c}_j^{(t)})m_j(\mathbf{c}_i^{(t)})}{m_i(\mathbf{c}_i^{(t)})m_j(\mathbf{c}_j^{(t)})}\right\}.$$

5. Record only the cold chain, m^1, for inferential purposes.

The key advantage to MCMCMC is that chains that get stuck in nonoptimal maxima will eventually get swapped out to some other, presumably more free, state. A notable disadvantage, though, is the need to possibly run many parallel chains for problems with highly complex targets.

2.4 *Simulated Tempering and Tempered Transitions*

Marinari and Parisi (1992) and (independently) Geyer and Thompson (1995) propose an alternative algorithm called *simulated tempering*, which reduces the MCMCMC algorithm to a single chain. Essentially the temperature itself becomes a random variable so the system can heat *and* cool as time proceeds. Why would one want do this in a simulated annealing process? Now elderly chains can still avoid being trapped at local maxima by getting more general Metropolis-Hastings candidate positions. That is, step 1 of the simulated annealing algorithm above is replaced with:

1a. Generate β from some distribution of temperature, $f(\beta)$.
1b. Generate $\mathbf{c}_j' \sim \left[\mathbf{c}_j^{(t')} + \mathcal{N}(0, \sigma^2)\right]$.

The number of $f(\beta)$ choices is obviously vast, but this decision can be simplified by using a discrete distribution that resembles some desired, but not implemented, cooling schedule. This algorithm can also be thought of as an augmented sampler in the context of Tanner and Wong (1987).

Neal (1996) builds on simulated tempering with *tempered transitions* to heat up the posterior distribution in place so that a random walk can move more freely, but also to preserve the detailed balance equation at each step.[1] See also Celeux et al. (2000) for an application to mixture distributions, and Liu and Sabatti (1999) for the "simulated simpering" variant. The basic idea is to "ladder" up and down in heat at each time t with random walk steps. Each ladder step specifies a (nonnormalized) stationary distribution defined on the same state space but at progressively hotter temperatures going up. Finally the last (bottom) ladder value is accepted or discarded with a Metropolis decision. This process is summarized by:

1. τ_1 is the target joint density.
2. β_i is the temperature value at the ith ladder step.
3. Tempered transitions: define a sequence of candidate densities m_i, $i = 1, \ldots, N$, where as i increases the m_i get "flatter" going up the ladder, then again more peaked going down the ladder.
4. Parameterize: $m_i = m^{1/\beta_i}$.
5. where: $1 < \beta_1 < \beta_2 < \cdots < \beta_{N-1} < \beta_N$.
6. Then: $\beta_N > \beta_{N+1} > \cdots > \beta_{2N-2} > \beta_{2N-1} > 1$.
7. Starting from the original candidate m, at each step we cycle through the m_i as follows:
   a. If we let MC($\mathbf{c}$,m) denote an MCMC kernel with position $\mathbf{c}$ and stationary distribution m,
   b. then we use the following transitions starting at iteration t:

$$\begin{aligned}
&\textbf{step 0:} && \mathbf{c}_{1,0}' \sim \mathrm{MC}(\mathbf{c}_1^{(t)}, \tau_1)\\
&\textbf{step 1:} && \mathbf{c}_{1,1}' \sim \mathrm{MC}(\mathbf{c}_{1,0}', m_1)\\
&\quad\vdots\\
&\textbf{step N:} && \mathbf{c}_{1,N}' \sim \mathrm{MC}(\mathbf{c}_{1,N-1}', m_N)\\
&\textbf{step N+1:} && \mathbf{c}_{1,N+1}' \sim \mathrm{MC}(\mathbf{c}_{1,N+1}', m_{N-1})\\
&\quad\vdots\\
&\textbf{step 2N-1:} && \mathbf{c}_{1,2N-1}' \sim \mathrm{MC}(\mathbf{c}_{1,2N-2}', m_1).
\end{aligned}$$

   c. The sequence of $\mathbf{c}_1$ values is then input into a final Metropolis-Hastings acceptance step, accepting $\mathbf{c}_{1,2N-1}'$ as $\mathbf{c}_1^{(t+1)}$ with probability:

$$\min\left\{1, \frac{m_1(\mathbf{c}_1^{(t)})}{\tau_1(\mathbf{c}_1^{(t)})} \cdots \frac{m_N(\mathbf{c}_{1,N-1}')}{m_{N-1}(\mathbf{c}_{1,N-1}')} \cdots \frac{\tau_1(\mathbf{c}_{1,2N-1}')}{m_1(\mathbf{c}_{1,2N-1}')}\right\},$$

   which preserves the detailed balance condition.

[1] For the posterior $\pi(\theta)$, define $MC(\theta',\theta)$ as the kernel of a MCMC algorithm going from θ to θ'. This Markov chain satisfies the detailed balance equation if $MC(\theta',\theta)\pi(\theta') = MC(\theta,\theta')\pi(\theta)$ (also called the *reversibility condition*).

To look at this in a slightly different way, we can also substitute the β_i parameterization back in. Now accept $\mathbf{c}_{1,2N-1}^{(t)}$ as $\mathbf{c}_1^{(t+1)}$ with probability:

$$\min\left\{1, \left(\frac{m^{1/\beta_1}(\mathbf{c}_{1,0}^{(t)})}{m^1(\mathbf{c}_{1,0}^{(t)})}\right)\left(\frac{m^{1/\beta_2}(\mathbf{c}_{1,1}^{(t)})}{m^{1/\beta_1}(\mathbf{c}_{1,1}^{(t)})}\right)\left(\frac{m^{1/\beta_3}(\mathbf{c}_{1,2}^{(t)})}{m^{1/\beta_2}(\mathbf{c}_{1,2}^{(t)})}\right)\right.$$
$$\cdots\left(\frac{m^{1/\beta_{N-1}}(\mathbf{c}_{1,N-2}^{(t)})}{m^{1/\beta_{N-2}}(\mathbf{c}_{1,N-2}^{(t)})}\right)\left(\frac{m^{1/\beta_N}(\mathbf{c}_{1,N-1}^{(t)})}{m^{1/\beta_{N-1}}(\mathbf{c}_{1,N-1}^{(t)})}\right)\left(\frac{m^{1/\beta_{N+1}}(\mathbf{c}_{1,N}^{(t)})}{m^{1/\beta_N}(\mathbf{c}_{1,N}^{(t)})}\right)$$
$$\left.\cdots\left(\frac{m^{1/\beta_2}(\mathbf{c}_{1,2N-3}^{(t)})}{m^{1/\beta_3}(\mathbf{c}_{1,2N-3}^{(t)})}\right)\left(\frac{m^{1/\beta_1}(\mathbf{c}_{1,2N-2}^{(t)})}{m^{1/\beta_2}(\mathbf{c}_{1,2N-2}^{(t)})}\right)\left(\frac{m^1(\mathbf{c}_{1,2N-1}^{(t)})}{m^{1/\beta_1}(\mathbf{c}_{1,2N-1}^{(t)})}\right)\right\}.$$

Note that $\mathbf{c}_{1,0}^{(t)} = \mathbf{c}_1^{(t)}$. We can also add a weighting function within each term above: $\frac{w(\mathbf{c}_{1,0}^{(t)})}{w(\mathbf{c}_{1,2N-1}^{(t)})}$ (sometimes called a *pseudo-prior* in this context). The original stationary distribution of the Markov chain is maintained as long as the m_i satisfy a detailed balance condition, which is given in Neal (1996) with proof.

This sequence of transitions allows excellent exploration of the parameter space, as the density m_N is typically chosen as very "hot," for example, uniform on the entire space. Setting $\beta_i - \beta_{i+1}$ as small gives higher acceptance rates but poorer mixing. Conversely a large difference between m and $m^{1/\beta}$ is good for mixing around the space but may lead to inordinately high rejection rates. Both criteria can be satisfied with taller ladders (i.e., more steps and a higher maximum temperature).

2.5 *Comparison of Algorithms*

It is interesting to compare these approaches with a deliberately "ugly" example objective function on $[-1,1]^2$:

$$f(x,y) = \text{abs}((x\sin(20y-90) - y\cos(20x+45))^3 a\cos(\sin(90y+42)x)$$
$$+ (x\cos(10y+10) - y\sin(10x+15))^2 a\cos(\cos(10x+24)y)).$$

This function is displayed in Fig. 1. While the vast majority of posterior forms will not exhibit such challenging characteristics, it still serves as a benchmark for algorithm performance. Furthermore, increasingly complex model specifications are much more prevalent in recent Bayesian and non-Bayesian work in political science, leading to potentially similar forms.

We now apply a regular random walk Metropolis-Hastings algorithm, simulated annealing, and tempered transitions to this function. Each Markov chain is run for 5000 iterations (an insufficient but illustrative period), and the chain visits are given in Fig. 2. Such a number of iterations is revealing here because, while all of these algorithms are ergodic and will therefore eventually explore the full target form, our concern is with the rate at which they do so. Specifically, do the algorithms differ in the efficiency by which they mix through the space?

It is clear that the standard algorithm fails during this period to break out of the diagonal and spends the bulk of its time visiting two of the four corners where large modes exist. The simulated annealing algorithm appears to be working much better but still fails to

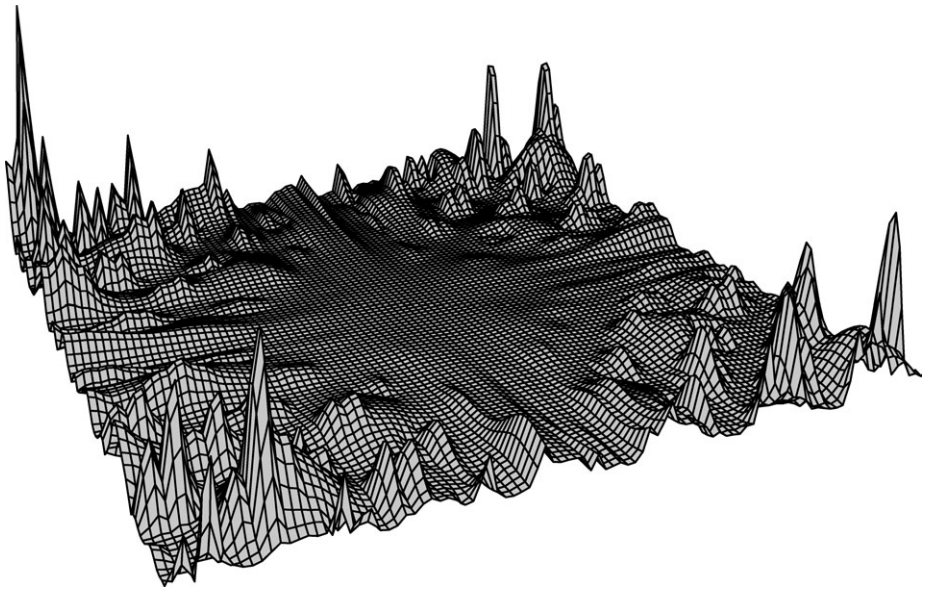

Fig. 1 A highly multimodal surface.

break out of the diagonal area. Conversely, the algorithm based on tempered transitions manages to explore the full state space, even in this small number of iterations. So the leverage that we gain from using the tempered transitions is greater assurance that we have mixed through the target form in a fixed amount of time.

3 Dynamic Tempered Transitions

We extend here the idea of tempered transitions to account for current placement of the chain in the state space. Our objective is to escape the necessary trade-offs in ladder height (maximum heating level) and spacing between rungs (number of steps). When the area around the chain is highly irregular, it is better to have a lower (cooler maximum temperature) ladder in order to get to the top of the mode and fully explore this area. When the area around the chain is smooth, it is better to have a longer (hotter maximum temperature) ladder in order to avoid being trapped in the low-density region. Also, setting the number of rungs for a given ladder height as too small can reduce the acceptance rate because high-quality candidates may not be offered. Conversely, settings the number of runs as too large can also reduce the acceptance rate because of the product in the Metropolis-Hastings decision step.

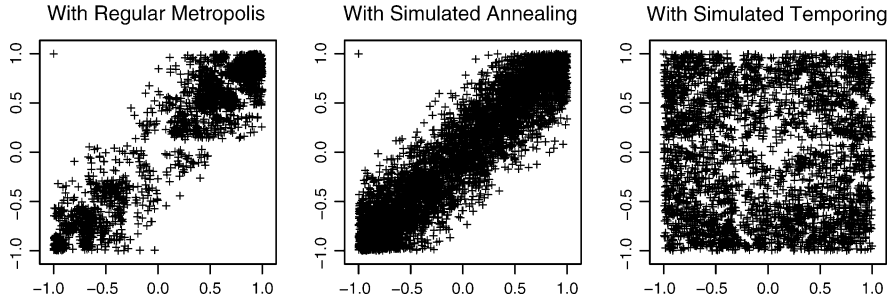

Fig. 2 A comparison of travels.

Our general strategy is to specify a distribution of ladders all having the same number of rungs but differing heights (differing maximum temperatures). We observe the multidimensional curvature at the current Markov chain location and specify a greater probability of selecting a cooler ladder when this curvature is high, and a greater probability of selecting a hotter ladder when the curvature is low. The number of rungs is essentially a nuisance parameter that can be fixed at the beginning of the chain or tuned during the early runs by comparing acceptance probabilities. The actual goal is not to determine the exact optimal number of rungs but to pick a reasonable number that does not overly affect the acceptance probabilities.

The key challenge is that by making the behavior of the Markov chain adjust to its surroundings (i.e., conditional on $\mathbf{c}$), we run the risk of creating a nonhomogeneous Markov chain and also losing the detailed balance equation. This would then deny us the ability to assert ergodicity of the chain. We can solve this problem by taking advantage of the structure of the Metropolis algorithm.

Let $f(\mathbf{c})$ be the stationary distribution (objective function), let $g(\mathbf{c}' \mid \mathbf{c})$ be a candidate distribution, and let $MC(\mathbf{c},\mathbf{c}')$ be the associated transition kernel. By the construction of the Metropolis algorithm, $MC(\mathbf{c},\mathbf{c}')$ is given by

$$MC(\mathbf{c},\mathbf{c}') = \min\left\{\frac{f(\mathbf{c}')g(\mathbf{c} \mid \mathbf{c}')}{f(\mathbf{c})g(\mathbf{c}' \mid \mathbf{c})}, 1\right\} g(\mathbf{c}' \mid \mathbf{c}) + (1 - r(\mathbf{c}))\delta_{\mathbf{c}}(\mathbf{c}'),$$

where

$$r(\mathbf{c}) = \int \min\left\{\frac{f(\mathbf{c}')g(\mathbf{c} \mid \mathbf{c}')}{f(\mathbf{c})g(\mathbf{c}' \mid \mathbf{c})}, 1\right\} g(\mathbf{c}' \mid \mathbf{c}) d\mathbf{c}'$$

and $\delta_{\mathbf{c}}(\mathbf{c}') = 1$ if $\mathbf{c} = \mathbf{c}'$ and zero otherwise. The kernel $MC(\mathbf{c},\mathbf{c}')$ now satisfies detailed balance with $f(\mathbf{c})$ as the stationary distribution (exactly from Robert and Casella 1999, Theorem 6.2.3).

Now for each $\mathbf{c}$, let $\rho(\lambda \mid \mathbf{c})$ be a probability distribution, that is, $\rho(\lambda \mid \mathbf{c}) \geq 0$ and $\int \rho(\lambda \mid \mathbf{c}) d\lambda = 1$. Here we are considering λ to be continuous, to be general, but usually λ will be discrete. Our candidate distribution is

$$g^*_\lambda(\mathbf{c}' \mid \mathbf{c}) = \rho(\lambda \mid \mathbf{c}) w_\lambda(\mathrm{c}, \mathrm{c}')$$

and forms the Metropolis kernel based on $g^*_\lambda(\mathbf{c}' \mid \mathbf{c})$ and $f(\mathbf{c})$ with the distribution of ladders $w_\lambda(\mathrm{c},\mathrm{c}')$ as in Neal (1996). By construction, detailed balance is satisfied and we have an ergodic Markov chain.

As an example, suppose that there are $i = 1, \ldots, k$ ladders, and as i increases the ladders get hotter. If $|f''(\mathbf{c})|$ is big (so we are near a mode) we might want to favor the cooler ladders. To do this we can take $\rho(\lambda \mid \mathbf{c})$ to be a binomial mass function with k trials and success probability $p(\mathbf{c})$, where

$$\operatorname{logit} p(\mathbf{c}) = a - b\,|f''(\mathbf{c})|, \quad b > 0.$$

Big values of $|f''(\mathbf{c})|$ would therefore result in small $p(\mathbf{c})$, which would favor the smaller values of i and the cooler ladders. This way, on average, we spend sufficient time exploring the modal area. Eventually, since there is always a positive probability of getting

a hot ladder, the chain eventually escapes from every mode. In flat areas, since hot ladders are more probable, this happens on average sooner. At http://psblade.ucdavis.edu we provide an easy to use (and modified) version of this algorithm written in R.

4 Motivating Example: Probabilistic Voting

We are centrally concerned in this example with the analysis of elections in spatial models, as studied by Coughlin and Nitzan (1981), Coughlin (1982, 1992), Enelow and Hinich (1984), Ledyard (1984), Ordeshook (1986), de Palma et al. (1990), Enelow et al. (1993), and Hinich and Munger (1994), where the focus is on *probabilistic* voting decisions. Important early works here include Luce and Raiffa (1957), Hinich et al. (1972, 1973), and Hinich (1977). Probabilistic voting models account for uncertain and unmeasured factors in the process by including a random, probabilistic element to each voter's preference calculation. The key idea is to replace a deterministic utility-maximizing decision with a continuously measured probabilistic calculation based on distance as well, but allowing votes for alternatives with lower expected utility with lower probability. The contrasts between probabilistic and deterministic voting are summarized in Calvert (1986), Mueller (1989), and Burden (1997). Essentially probabilistic voting means that voters have disturbance terms, which means that they take into account some omitted variables that are not in the calculus of the candidates (Erikson and Romero 1990).

This example is based on Enelow and Hinich (1984), and the theoretical framework is a variation developed by Lin et al. (1999) extended here to include model components that these authors were unable to estimate empirically due to algorithmic and computational limitations. We take as our theoretical/formal starting point (as do many others) the multicandidate probabilistic voting construct directly from Shepsle (1972) and Enelow and Hinich (1984). The setup begins with J candidates, N voters, and a compact, convex K-dimensional Euclidean issue space $\mathbb{S}^K$. We use the proximity voting model without necessarily denigrating the directional voting, since both have theoretical insights (see Lewis and King's [2000] or Cho and Endersby's [2003] analysis of this literature along with the strong adherents: Westholm [1997] for proximity and MacDonald et al. [1998] for direction). The candidates each pick a point in $\mathbb{S}^K$ designed to maximize their expected votes (interestingly, maximizing expected votes is different from maximizing the probability of victory; see Patty [2002]). Conversely, Kollman et al. (1992a) point out that sometimes parties or candidates prefer to be *ideological* in that they will constrain their range of allowable positions even in the presence of higher expected returns elsewhere. Define the position of candidate j ($j = 1, \ldots, J$) to be the K-length vector: $\mathbf{c}_j = [C_{j1}, C_{j2}, \ldots, C_{jK}]$. Voter i's *ideal point* in $\mathbb{S}^K$ is the K-length vector: $\mathbf{v}_i = [V_{i1}, V_{i2}, \ldots, V_{iK}]$. The voter is assumed to vote for the candidate who has a K-dimensional position the closest to this voter's ideal point; this is *sincere voting*.

We can also define the error contribution that candidate j produces by making vague statements or avoiding specific issues in general, $\mathbf{E}_j$. This component can be thought of as an error term in the statistical sense, since it partly obscures the single point from the Euclidian distance that would result from definitive statements (measurement is performed by the observer, but measurement error emanates from the target). In a more general sense, the error term can also come from other sources such as issues unrelated to specific policy stands, as well as policy considerations that are left unaddressed by the explicit terms in the model.

Following Lin et al. (1999, p. 62) for simplicity in the choices of our integral calculations (below) and the subsequent computational process, we assume that the error terms are individual observations from a zero-mean random variable so that the sum of

these errors over J candidates to N voters equals the sum of errors over the N voters to J candidates. While this is certainly a substantial simplifying assumption, since it implies candidate uniformity to voters, the subsequent model still produces highly variable, multimodal objective functions that trouble conventional descriptive algorithms.

The utility of candidate j to voter i is the negative (vector) distance between $\mathbf{c}_j$ and $\mathbf{v}_i$, plus the error term:

$$\mathbf{U}_{ij} = \mathbf{E}_j - \mathbf{D}_{ij} = \mathbf{E}_j - ||\mathbf{v}_i - \mathbf{c}_j||, \tag{1}$$

where $||.||$ denotes the vector K norm. The factor $\mathbf{D}_{ij}$ is the usual Euclidean distance here, but it is also common to stipulate squared Euclidean distance ($||\mathbf{v}_i - \mathbf{c}_j||^2$), salience-weighted squared Euclidean distance ($(\mathbf{v}_i - \mathbf{c}_j)'\Omega(\mathbf{v}_i - \mathbf{c}_j)$, where Ω is a symmetric, positive definite matrix), and so-called city-block distance: the summed absolute differences. Sometimes these different utility measures provide different results (Enelow et al. 1988), but often the results are quite similar (Lewis and King 2000).

Substantive covariates can be additively included to reflect various voter characteristics unrelated to issue distance (Erikson and Romero 1990). Adams and Merrill (2001; see also Merrill and Adams 2002), for instance, specify a salience-weighted (a) squared Euclidean distance measure that also includes salience-weighted ($\mathbf{B}$) individual nonpolicy characteristics ($\mathbf{t}_i$) such as demographics and social characteristics $\mathbf{U}_{ijk} = -a(\mathbf{v}_i - \mathbf{c}_j)^2 + \mathbf{B}\mathbf{t}_i$ (see also Enelow and Hinich [1982] and Yang [1995] for inclusion of so-called nonpolicy terms). The associated generalization of (1) is $\mathbf{U}_{ij} = \mathbf{E}_j - f(\mathbf{v}_i - \mathbf{c}_j) + g_{ij}(\boldsymbol{\theta} \mid \psi)$, where $\mathbf{E}_j$ is a normal error, f is a distance function, and g_{ij} is a covariate function that affects the nonpolicy value of candidate j to voter i. The $\boldsymbol{\theta} \mid \psi$ 'nonpolicy' covariates include demographic, social, and political variables on individual voters that are distinct from issue distance: income, education, party affiliation, family characteristics, religion, and others, which can be hierarchically conditioned on other parameters (ψ).

The standard assumption is that the error term ($\mathbf{E}_j$) is independent across candidates and has zero mean and thus can be negative (utility decreasing) or positive (utility increasing). Furthermore, we specify the variance of each individual candidate σ_j^2 to reflect this vagueness (Campbell 1983). Thus larger values of σ imply that the candidate is less clear to voters about issue position. It turns out that the assumptions underlying assignment of the J values of σ are critical to the quality of the resulting model.

This setup allows the direct comparison of two candidates for any given voter. That is, voter i prefers candidate j over candidate ℓ if her utility for j exceeds her utility for ℓ:

$$\mathbf{U}_{ij} - \mathbf{U}_{i\ell} = \mathbf{E}_j - \mathbf{D}_{ij} - \mathbf{E}_\ell + \mathbf{D}_{i\ell} > 0,$$

where Enelow and Hinich further define $\mathbf{E}_{j\ell} = \mathbf{E}_\ell - \mathbf{E}_j$ and $\mathbf{D}_{ij,\ell} = -\mathbf{D}_{ij} + \mathbf{D}_{i\ell}$. Since $\mathbf{E}_{j\ell}$ is the sum of two zero-mean random variables, it too has zero mean with additive variance $Var(\mathbf{E}_{j\ell}) = \sigma_j^2 + \sigma_\ell^2$. Of course voters are assumed to be comparing *all* candidates, so for j as the "baseline" comparison candidate, the $(J - 1)$-length error vector is usually treated as multivariate normal:

$$\mathbf{e}_j = [\mathbf{E}_{j1}, \ldots, \mathbf{E}_{j(j-1)}, \mathbf{E}_{j,(j+1)}, \ldots, \mathbf{E}_{jJ}] \sim \phi(\mathbf{0}, \Delta_j).$$

If we now collect the K-dimensional $J - 1$ distance cross-candidate comparisons to candidate j into a single vector, $\mathbf{d}_j = [\mathbf{D}_{ij,1}, \ldots, \mathbf{D}_{ij,(j-1)}, \mathbf{D}_{ij,(j+1)}, \ldots, \mathbf{D}_{ij,J}]$, then the

probability that voter i votes for candidate j is the CDF (cumulative distribution function) of the multivariate normal at the (vector-valued) point $\mathbf{d}_j$:

$$\begin{aligned} P(i,j) &= P(\mathbf{E}_{j\ell} < \mathbf{D}_{ij,\ell}, \ell = 1,\ldots,j-1,j+1,\ldots,J) \\ &= \int_{-\infty}^{\mathbf{D}_{ij,1}} \cdots \int_{-\infty}^{\mathbf{D}_{ij,j-1}} \int_{-\infty}^{\mathbf{D}_{ij,j+1}} \cdots \int_{-\infty}^{\mathbf{D}_{ij,J}} \phi(\mathbf{0},\Delta_j) d\mathbf{E}_{j,1}\cdots d\mathbf{E}_{j,j-1} d\mathbf{E}_{j,j+1}\cdots d\mathbf{E}_{j,J}, \end{aligned} \tag{2}$$

where we have used the fact that $\mathbf{E}_{j\ell}$ is a symmetric mean-zero random variable to reverse the inequality. This setup makes it easy to calculate the expected vote totals for each candidate:

$$EV_j(\mathbf{c}) = \sum_{i=1}^{N} P(i,j)$$

along with the associated expected vote proportion:

$$\tau(\mathbf{c} \mid \mathbf{v}) = \sum_{i=1}^{N} P(i,j)/N.$$

4.1 *Illustration*

With this simple setup it is possible to calculate the expected electoral fortunes of a given candidate for K-dimensional positions through the issue space where the other candidates are assumed to be fixed at known positions. This can be either a static situation such as a proposal given to a legislature or one iteration of a more complex game. Note, however, that there is no equilibrium point in this basic setup if *all* the candidates are allowed to simultaneously make the calculation done by one here.

Consider, for simplicity, a three-candidate race with 100 voters over a standardized two-dimensional issue metric, where we evaluate the expected number of votes for candidate *1* taking all possible issue positions. The voter ideal points are drawn from a mixture of beta densities to reflect some division of preferences across two roughly defined groups. Candidates 2 and 3 have the moderate but opposing positions of [0.4, 0.6] and [0.6, 0.4], respectively, and it is assumed that $\mathbf{e}_j \sim \mathcal{N}(0,\ 0.02)\ \forall j$. The resulting expected vote for candidate *1* is given in Fig. 3, where the left-hand panel gives the three-dimensional blanket of expected value and the right-hand panel shows a corresponding contour plot with the 100 voter ideal points marked. In fact, candidate *1* can maximize her fortunes by picking an issue position around one of two dominant modes, but she apparently cannot achieve a simple majority at any position in the issue space (note also that there are no voters located directly at the two modes). Even this simple demonstration produces a relatively intricate expected vote form, and adding more realistic assumptions can substantially complicate the task of strategic mode finding.

4.2 *In Pursuit of Equilibria*

A preferred candidate point for these models considers the full set of other candidate positions and determines where this candidate can no longer improve her vote share by

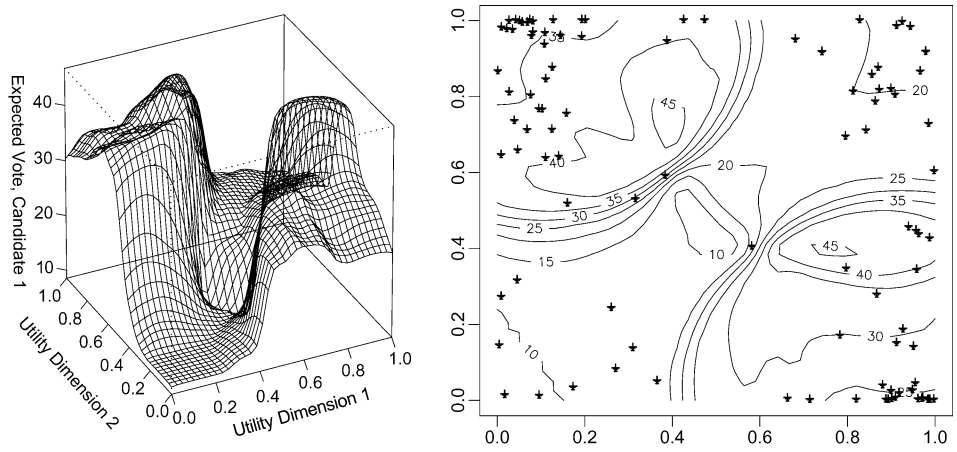

Fig. 3 Vote expectations for candidate 1.

moving in issue space. Lin et al. (1999, Appendix 1) show that for this candidate's vote function, the gradient of $EV_j(\mathbf{c})$ is given by

$$\frac{\partial}{\partial \mathbf{c}_j} EV_j(\mathbf{c}) = -\frac{\mathbf{D}_{ij}}{\partial \mathbf{c}_j} \sum_{i=1}^{N} \left[\underbrace{\frac{\partial \phi}{\partial \mathbf{D}_{ij,1}} + \ldots + \frac{\partial \phi}{\partial \mathbf{D}_{ij,(j-1)}} + \frac{\partial \phi}{\partial \mathbf{D}_{ij,(j+1)}} + \ldots + \frac{\partial \phi}{\partial \mathbf{D}_{ij,J}}}_{g(\mathbf{d}_j|\Delta_j)} \right],$$

with the associated Hessian:

$$\frac{\partial^2}{\partial \mathbf{c}_j \partial \mathbf{c}_j'} EV_j(\mathbf{c}) = -\sum_{i=1}^{N} \left[g(\mathbf{d}_j \,|\, \Delta_j) \frac{\partial^2 \mathbf{D}_{ij}}{\partial \mathbf{c}_j \partial \mathbf{c}_j'} - \underbrace{\left(\frac{\partial^2 \phi}{\partial \mathbf{D}_{ij,1}^2} + \ldots + \frac{\partial^2 \phi}{\partial \mathbf{D}_{ij,(j-1)}^2} + \frac{\partial^2 \phi}{\partial \mathbf{D}_{ij,(j+1)}^2} + \ldots + \frac{\partial^2 \phi}{\partial \mathbf{D}_{ij,J}^2} \right)}_{h(\mathbf{d}_j|\Delta_j)} \frac{\mathbf{D}_{ij}}{\partial \mathbf{c}_j} \frac{\mathbf{D}_{ij}}{\partial \mathbf{c}_j'} \right].$$

A vote-maximizing point for candidate j exists when there is a maxima that meets the concavity condition: the gradient is found to be zero (the first-order condition) and the Hessian is positive semidefinite (the second-order condition). Critically, as overall candidate error levels increase ($\Delta_j \uparrow$), the form of $EV_j(\mathbf{c})$ "flattens out" and tends toward unimodality [i.e., $h(\mathbf{d}_j \,|\, \Delta_j)/g(\mathbf{d}_j \,|\, \Delta_j)$] gets smaller). This finding from the standard Enelow-Hinich setup means that the greater the random term, the easier it is to find equilibria. However, in situations where one does not have the "luxury" of assuming relatively big errors in the model that smooth out the objective function (perhaps due to high-quality information flow), then the functional forms can be substantially more complex and beyond the abilities of more primitive computing solutions.

Partially successful computing solutions to such problems exist. Herzberg and Wilson (1988) were able to determine strategic versus sincere equilibria in a highly controlled experimental setting. Adams and Merrill (2001) are able to find multiple equilibria as long as the

model is specified with a conditional logit function for voter preference and issue distance is assumed to be a squared Euclidean measure. Kollman et al. (1992b) use simulated annealing to find multiple modes in a model institutional responsiveness to voter heterogeneity. The Lin et al. (1999) solution can be characterized as requiring that the stochastic component dominate the policy component in order to obtain equilibria, which contradicts a considerable amount of empirical work on voting behavior (Carmines and Stimson 1980; Bartels 1988; Aldrich et al. 1989; Goren 1997; Alvarez et al. 2000). Others have simplified the policy space down to categorical evaluations (Iverson 1994), basic ideology space with varying levels of information (Martinelli 2001) or designed simple experiments with students as voting subjects (McKelvey and Ordeshook 1985; Morton 1993).

This is where we pick up: the development of a comprehensive estimation method for arbitrarily complex voting questions without necessitating analytical derivation as described above, particularly since some forms are intractable. Such problems arise naturally with large numbers of voters or candidates, bidirectional uncertainty, and large numbers of simultaneously considered dimensions. The posterior surfaces that arise generally contain a large number of nonoptimal modes and relatively complex shapes in general. Overcoming this challenge allows us to ask a number of substantive questions about voting, but in more realistic settings. Our algorithm is undeterred by estimation complexities that haunt this literature. For instance, complex forms that arise in the estimation of both underlying voter distributions (rather than the fixed positions given above) and the location of candidates' vote-maximizing positions are easily handled once the full objective function is described by a theoretical model.

4.3 *Application to a Real Dataset*

Purely as a means of demonstrating the capabilities of this new Bayesian stochastic simulation procedure, we apply it to the American National Election Study (ANES) of the 2000 presidential election, with 1462 potential voters surveyed prior to the election. The point is not to reanalyze this election result, but rather to show the ease with which the dynamic algorithm handles large datasets and high dimensions. Specifically, can we use the Markov chain to mix efficiently through an 11-dimensional parameter space in relatively few iterations in order to find a global maxima?

First, we look at 10 policy dimensions in which respondents place themselves, to produce an ideal point vector: $\mathbf{v}_i = [V_{i1}, V_{i2}, \ldots, V_{iK}]$. The studied policy issues given by the nominal scales (wording shortened from survey) are:

- political ideology
- preference on increased government spending
- preference on increased defense spending
- government should help generate jobs
- government should help blacks economically
- support for abortion rights
- environment versus industrial development
- support for gun ownership rights
- central role of women in society
- increased/decreased regulation

The multidimensional positions for Gore and Bush are modal evaluations from respondents across 10 dimensions:

$$\mathbf{c}_{Gore} = [C_{Gore,1}, C_{Gore,2}, \ldots, C_{Gore,10}]$$
$$\mathbf{c}_{Bush} = [C_{Bush,1}, C_{Bush,2}, \ldots, C_{Bush,10}].$$

This has the advantage of using *voter* perceptual placement of the candidates rather than some other measure such as interest group rankings, which would necessitate worrying about differences in information or campaign effects. Our question in this simple example is, can we find a third (purely strategic and nonideological) candidate position that beats both Bush and Gore on these policy issues? Specifically,

$$\tau(\mathbf{c}_{new} \mid \mathbf{v}) > \tau(\mathbf{c}_{Gore} \mid \mathbf{v}) \quad \text{and} \quad \tau(\mathbf{c}_{Bush} \mid \mathbf{v}).$$

Or, as Erikson and Romero (1990) put it in a broader context: "The general quest is for a location candidate policy equilibrium—a location in policy space to which knowledgeable candidates should gravitate if they want to win elections." In fact, our test builds slightly on the Erikson and Romero example. They look at fixing (the elder) Bush's multidimensional policy position and seeking the best position for an accordingly strategic Dukakis candidate. We are somewhat more complicated in that both Bush's and Gore's positions are fixed and a hypothetical third candidate is evaluated strategically.

Our dynamic tempering MCMC algorithm traverses through this 11-dimensional posterior issue space where the objective function is determined by sums from three candidates and 1462 voters. After a reasonable burn-in period of 10,000 iterations we record the following 50,000 iterations. Standard diagnostics give us no reason for concern about convergence (evaluations included Fig. 4 with a cumulative sum plot over each dimension and a superimposed trace plot).

What does this mean electorally? We compare the modal positions assigned by respondents for $\mathbf{c}_{Gore}$ and $\mathbf{c}_{Bush}$ to the posterior mean from the model. Note that this position is different than the data mean, which is also provided in Table 1. We recognize that the mean as a summary overstates the level of measurement here, but it remains useful for comparing candidates and may not be far from even spaced.

Now set the hypothetical "new" candidate at the posterior mean as a strategic choice and use $EV_{new}(\mathbf{c}) = \sum_{i=1}^{N} P(i, new)$ to produce their expected fortunes at this position. To obtain the $\tau(j)$ probability, we simply apply importance sampling rather than analytical calculation of the integral value. It is worth noting that once the MCMC samples are collected, we can easily calculate many different empirical summaries in addition to those used here. The probabilities for the two actual candidates at their modes and the hypothetical candidate are:

$$\tau(Gore \mid \mathbf{v}) = 0.34758, \qquad \tau(Bush \mid \mathbf{v}) = 0.17477, \qquad \tau(Candidate \mid \mathbf{v}) = 0.47100.$$

Thus our hypothetical candidate would win under plurality voting. Obviously there are greater complications here, including the electoral college, which was pivotal in this election. Nonetheless, it is possible to construct a purely strategic candidate position in this case that garners higher expected vote probability than the actual candidates, assuming the perceived positions remain fixed. This result, of course, ignores personality, retrospective voting, major party bias, and a whole host of other issues. We have shown, however, that a combined large number of voters, candidates, and issues is not a serious deterrent for the algorithm. Furthermore, any number of additional enhancements, such as replacing voter ideal points with voter distributions to reflect uncertainty, adding candidate restrictions, and iterative game structures, will not deter the dynamic algorithm. As long as the target

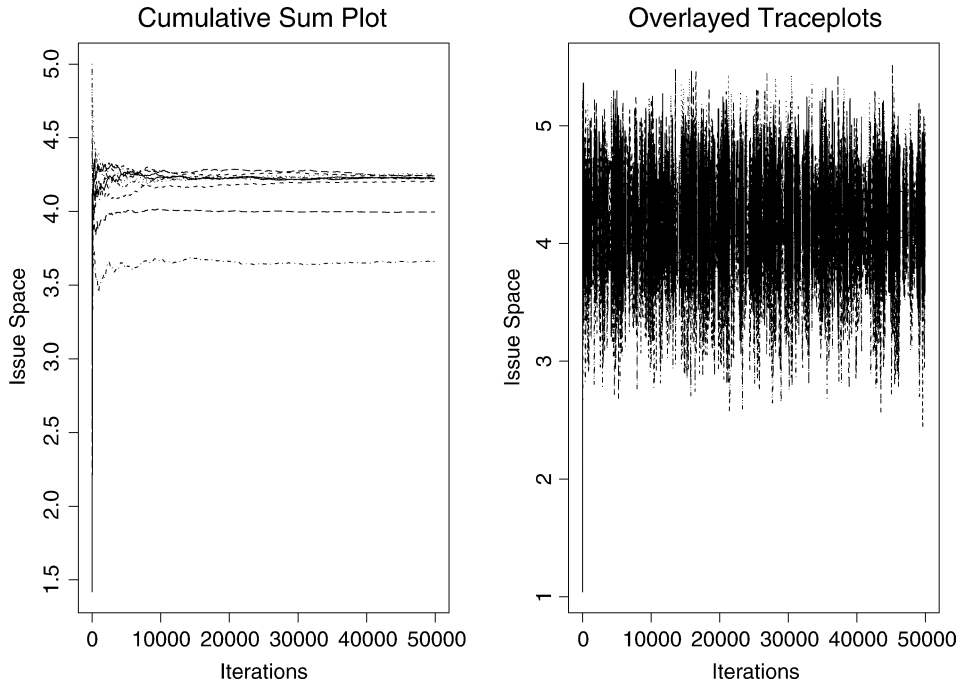

Fig. 4 Two diagnostics.

can be mathematically described as a joint function, our approach will work, despite the complexity of this form, and we hope that scholars focused on such substantive questions will make use of this new tool.

5 Conclusion

The purpose of this paper was to introduce a new Bayesian stochastic simulation algorithm that performs well in the presence of multimodal posterior forms. As noted, these functional forms typically cause MCMC algorithms to get "stuck" in minor modes for extended periods of time, thus diminishing their ability to describe the full distribution of interest in reasonable periods. The result, unfortunately, is a biased summary estimate of the effect of interest.

The dynamic alternative specified here performs better in difficult circumstances than do standard Metropolis-Hastings setups because it can adjust to its current position. In

Table 1 Policy positions for real and hypothetical candidates

	Ideology	*Spending*	*Defense*	*Jobs*	*Blacks*	*Abortion*	*Environment*	*Guns*	*Gender*	*Regulation*
Bush										
mode	6	4	4	6	4	5	4	4	2	3
Gore										
mode	2	3	4	4	4	2	4	2	2	2
Data										
mean	4.3352	3.7770	4.1389	4.4234	4.5937	4.1115	3.6272	3.0116	2.6238	3.5903
Posterior										
mean	3.9423	4.2478	4.2384	4.2195	4.2230	4.2494	4.3388	4.2412	3.6322	4.2006

modifying or creating MCMC algorithms, it is essential to show that the well-known desirable properties leading to convergence are not destroyed. A Metropolis-Hastings algorithm that preserves the detailed balance equation is known to be ergodic, and ergodic Markov chains will eventually converge (Meyn and Tweedie 1994). Thus our ability to show that the dynamic tempered transitions algorithm preserves the detailed balance equation, despite conditioning on the present posterior characteristics, is a proof of eventual convergence on par with any standard MCMC procedure.

We also note that theoretical and empirical problems of interest to political scientists can readily lead to complex and multimodal posterior forms. MCMC tools represent an opportunity for empirical political scientists to estimate quantities of interest that were previously unobtainable. Yet standard algorithms are no less sensitive to such complexity than traditional mode-finding procedures. Thus we provide a practical solution to a vexing and common problem.[2]

References

Adams, James, and Samuel Merrill III. 2001. "A Theory of Spatial Competition with Biased Voters: Party Policies Viewed Temporally and Comparatively." *British Journal of Political Science* 31:121–158.

Aldrich, John H., John L. Sullivan, and Eugene Borgida. 1989. "Foreign Affairs and Issue Voting: Do Presidential Candidates 'Waltz Before A Blind Audience?' *American Political Science Review* 83:123–141.

Altman, Micah, Jeff Gill, and Michael McDonald. 2004. *Numerical Issues in Statistical Computing for the Social Scientist.* New York: John Wiley & Sons.

Alvarez, R. Michael, Jonathan Nagler, and Shaun Bowler. 2000. "Issues, Economics, and the Dynamics of Multiparty Elections: The 1997 British General Election." *American Political Science Review* 94:131–149.

Bartels, Larry. 1988. "Issue Voting under Uncertainty: An Empirical Test." *American Journal of Political Science* 30:709–728.

Burden, Barry C. 1997. "Deterministic and Probabilistic Voting Models." *American Journal of Political Science* 41:1150–1169.

Calvert, Randall L. 1986. *Models of Imperfect Information in Politics.* London: Harwood.

Campbell, James E. 1983. "Ambiguity in the Issue Positions of Presidential Elections: A Causal Analysis." *American Journal of Political Science* 27:284–293.

Carmines, Edward G., and James A. Stimson. 1980. "The Two Faces of Issue Voting." *American Political Science Review* 74:78–91.

Celeux, G., M. Hurn, and C. P. Robert. 2000. "Computational and Inferential Difficulties with Mixture Posterior Distributions." *Journal of the American Statistical Association* 95:957–970.

Cho, Sungdai, and James W. Endersby. 2003. "Issues, the Spatial Theory of Voting, and British General Elections: A Comparison of Proximity and Directional Models." *Public Choice* 114:275–293.

Coughlin, Peter J. 1982. "Pareto Optimality of Policy Proposals with Probabilistic Voting." *Public Choice* 39:427–433.

Coughlin, Peter J. 1992. *Probabilistic Voting Theory.* Cambridge: Cambridge University Press.

Coughlin, Peter J., and Samuel Nitzan. 1981. "Election Outcomes with Probabilistic Voting and Nash Social Welfare Maxima." *Journal of Public Economics* 15:113–122.

de Palma, A., G.-S. Hong, and J.-F. Thisse. 1990. "Equilibria in Multi-party Competition under Uncertainty." *Social Choice Welfare* 7:247–259.

Enelow, James, M., James W. Endersby, and Michael C. Munger. 1993. "A Revised Spatial Model of Elections: Theory and Evidence." In *Information, Participation, and Choice,* ed. Bernard Grofman. Ann Arbor: University of Michigan Press, pp. 125–140.

Enelow, James M., and Melvin J. Hinich. 1982. "Nonspatial Candidate Characteristics and Electoral Competition." *Journal of Politics* 44:115–130.

Enelow, James M., and Melvin J. Hinich. 1984. *The Spatial Theory of Voting.* Cambridge: Cambridge University Press.

Enelow, James M., Nancy R. Mendell, and Subha Ramesh. 1988. "A Comparison of Two Distance Metrics through Regression Diagnostics of a Model of Relative Candidate Evaluation." *Journal of Politics* 50:1057–1071.

[2]Code is available at our Web site: http://psblade.ucdavis.edu.

Erikson, Robert S., and David W. Romero. 1990. "Candidate Equilibrium and the Behavioral Model of the Vote." *American Political Science Review* 84:1103–1126.

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85:389–409.

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

Geyer, C., and E. Thompson. 1995. "Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference." *Journal of the American Statistical Association* 90:909–920.

Gill, Jeff, and Gary King. Forthcoming. "Alternatives to Model Respecification in Nonlinear Estimation." *Sociological Methods and Research*.

Goren, Paul. 1997. "Political Expertise and Issue Voting in Presidential Elections." *Political Research Quarterly* 50:387–412.

Herzberg, Roberta Q., and Rick K. Wilson. 1988. "Results on Sophisticated Voting in an Experimental Setting." *Journal of Politics* 50:471–486.

Hinich, Melvin. 1977. "Equilibrium in Spatial Voting: The Median Voter is an Artifact." *Journal of Economic Theory* 16:208–219.

Hinich, Melvin, John Ledyard, and Peter Ordeshook. 1972. "Nonvoting and the Existence of Equilibrium under Majority Rule." *Journal of Economic Theory* 14:144–153.

Hinich, Melvin, John Ledyard, and Peter Ordeshook. 1973. "A Theory of Electoral Equilibrium: A Spatial Analysis Based on the Theory of Games." *Journal of Politics* 35:154–193.

Hinich, Melvin, and Michael C. Munger. 1994. *Ideology and the Theory of Political Choice*. Ann Arbor: University of Michigan Press.

Iverson, Torben. 1994. "Political Leadership and Representation in West European Democracies: A Test of Three Models of Voting. *American Journal of Political Science* 38:45–74.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. "Optimization by simulated annealing." *Science* 220:671–680.

Kollman, Ken, John H. Miller, and Scott E. Page. 1992a. "Adaptive Parties in Spatial Elections." *American Political Science Review* 86:929–937.

Kollman, Ken, John H. Miller, and Scott E. Page. 1992b. "Political Institutions and Sorting in a Tiebout Model." *American Economic Review* 87:977–992.

Ledyard, J. O. 1984. "The Pure Theory of Large Two-Candidate Elections." *Public Choice* 44:7–41.

Lewis, Jeffrey, and Gary King. 2000. "No Evidence on Directional vs. Proximity Voting." *Political Analysis* 8:21–33.

Lin, Tse-Min, James M. Enelow, and Han Dorussen. 1999. "Equilibrium in Multicandidate Probabilistic Voting." *Public Choice* 98:59–82.

Liu, J. S., and C. Sabatti. 1999. "Simulatated Sintering: Markov Chain Monte Carlo with Spaces Varying Dimension." In *Bayesian Statistics*, eds. J. M. Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger. Oxford: Oxford University Press, pp. 389–414.

Luce, R. D., and H. Raiffa. 1957. *Games and Decisions.* New York: John Wiley & Sons.

Macdonald, Stuart Elaine, George Rabinowitz, and Olga Listhaug. 1998. "On Attempting to Rehabilitate the Proximity Model: Sometimes the Patient Just Can't be Helped." *Journal of Politics* 60:653–690.

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

Martinelli, César. 2001. "Elections with Privately Informed Parties and Voters." *Public Choice* 108:147–167.

McKelvey, Richard D., and Peter C. Ordeshook. 1985. "Sequential Elections with Limited Information." *American Journal of Political Science* 29:480–512.

Merrill, Samuel III, and James Adams. 2002. "Centrifugal Incentives in Multi-candidate Elections." *Journal of Theoretical Politics* 14:275–300.

Meyn, S. P., and R. L. Tweedie. 1994. "State-Dependent Criteria for Convergence of Markov Chains." *Annals of Applied Probability* 4:149–168.

Morton, Rebecca B. 1993. "Incomplete Information and Ideological Explanations of Platform Divergence." *American Political Science Review* 382–407.

Mueller, Dennis C. 1989. *Public Choice II.* Cambridge: Cambridge University Press.

Neal, R. 1996. "Sampling from Multimodal Distributions Using Tempered Transitions." *Statistics and Computing* 4:353–366.

Nelder, J. A., and R. W. M. Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A* 135:370–385.

Ordeshook, Peter C. 1986 *Game Theory and Political Theory.* Cambridge: Cambridge University Press.

Patty, John W. 2002. "Equivalence of Objectives in Two Candidate Elections." *Public Choice* 112:151–166.

Rasmussen, Carl Edward. 2003. "Gaussian Processes to Speed up Hybrid Monte Carlo for Expensive Integrals." In *Bayesian Statistics 7,* eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, and A. P. Dawid. Oxford: Oxford University Press.

Robert, C. P., and G. Casella. 1999. *Monte Carlo Statistical Methods.* New York: Springer-Verlag.

Shepsle, Kenneth A. 1972. "The Strategy of Ambiguity: Uncertainty and Electoral Competition." *American Political Science Review* 66:555–568.

Tanner, M. A., and W. H. Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Society* 82:528–550.

Westholm, Anders. 1997. "Distance Versus Direction: The Illusory Defeat of the Proximity Theory of Electoral Choice." *American Political Science Review* 91:865–885.

Yang, C. C. 1995. "Endogenous Tariff Formation under Representative Democracy: A Probabilistic Voting Model." *American Economic Review* 85:956–963.