Movies for the Visualization of MCMC Output

Nicole A. Lazar	Joseph B. Kadane
Department of Statistics	Department of Statistics
Carnegie Mellon University	Carnegie Mellon University
Pittsburgh, PA 15213	Pittsburgh, PA 15213
nlazar@stat.cmu.edu	kadane@stat.cmu.edu

Summary

Much work has been done on visualization methods for Markov Chain Monte Carlo analyses, although there are still many open questions. Particular difficulties arise when the researcher wishes to monitor several aspects of the chain's behavior jointly. We propose a movie, which is an extension of the basic trace plot used for a single parameter, as a dynamic way of assessing the progress of a chain and monitoring the joint posterior distribution. This idea is demonstrated on several examples, including a multiple changepoint problem with data from a corbelled tomb (*tholos*) in Stylos, Crete.

Key words: Archaeology, Multiple Changepoints, Reversible Jump MCMC, Sampling–Resampling, Visualization

1 Introduction

Markov Chain Monte Carlo (MCMC) simulation techniques have been a boon for modern statistics, in particular in the analysis of complex data structures from a Bayesian perspective. The methods, namely the Metropolis–Hastings algorithm and the Gibbs sampler, are general enough to handle a wide variety of situations, which has led to increased popularity and widespread use. This is especially true in the last twenty years or so, as computing power has increased to match the richness of the models now being analyzed.

A major theoretical issue surrounding the use of MCMC is examining the convergence of a chain. This relates both to drawing samples from the target, or stationary distribution, and to ensuring that the chain traverses the entire relevant parameter space in a reasonable amount of time, a property known as mixing. While the theory of Markov chains, on which this simulation technique is based, guarantees that convergence *will* occur, *i.e.*, that the stationary distribution will be reached, it offers no guidance as to how long this will take, nor how to assess when a given chain has started to sample from the distribution in question. Researchers have proposed various diagnostics, suitable for different implementations of the general methodology (for instance, diagnostics for chains run using the Gibbs sampler only, or for comparing multiple chains run in parallel from different starting points). These diagnostics are based on a variety of theoretical justifications, ranging from large–sample normal theory (Gelman and Rubin, 1992), to spectral analysis (Geweke, 1992), to graphical methods based on cumulative sum path plots (Yu and Mykland, 1998). Unfortunately, as pointed out by Cowles and Carlin (1996), all of the proposed methods can fail to detect departures from convergence, even for relatively simple problems. They therefore recommend trying multiple diagnostics on any given problem; most of the standard analyses can be performed using the Bayesian Output Analysis program, or BOA, available at `http://www.public-health.uiowa.edu/boa`. Less formally, it is also common to use graphical techniques such as the trace plot, which plots the marginal progressions of the model parameters over the course of the chain, to assess convergence and evaluate mixing.

Ordinary MCMC methods are applicable when the dimension of the problem is fixed, for example, analyzing a model with a fixed number of components in a mixture, or a fixed number of piecewise constants. If the dimension of the problem is itself taken to be a parameter to be estimated via its posterior distribution, more sophisticated algorithms are needed. There are a number of alternatives suggested in the literature for treating such a problem. Chib (Chib, 1995; Chib and Jeliazkov, 2000) suggests running separate chains conditional on the dimension of the parameter space, and separately computing the probability of each dimension. The reversible jump MCMC method (Green, 1995; Richardson and Green, 1997), gives a single chain sampling from the joint posterior on the unconditional space. As might be expected from the previous discussion, it is not at all clear how to evaluate the validity of the results from such simulations, nor are there generally agreed upon methods for summarizing and displaying the output of these chains.

For assessing multiple parameters jointly, or for monitoring the convergence of a reversible jump MCMC run, extensions of the existing methods are needed. Brooks and Giudici (2000), in the context of model choice based on reversible jump MCMC, propose monitoring the convergence of a parameter that has a coherent meaning throughout the run of the simulation. For example, in a mixture model problem, where the number of components is not known, the researcher might choose to assess the chain as a whole based on the convergence of the parameter that counts the number of components. The method builds on suggestions of Gelman and Rubin (1992) and Brooks and Gelman (1998) to run multiple chains and use the variability within and between chains to determine stationarity. However, the applicability of the technique to complicated chains depends on being able to find a parameter, or set of parameters, whose meaning is comparable across different models, which might not always be possible. Additionally, convergence of such a parameter does not necessarily imply convergence of the chain viewed multi–dimensionally.

This paper proposes a complementary approach, namely, an extension of the standard trace plot used to monitor unidimensional convergence, to a multidimensional, statistical movie, that allows the statistician to follow the progress of a chain over successive iterations and to monitor changes in many (perhaps all) of the parameters, including the dimension of the model where relevant, simultaneously. The outline of the rest of the article is as follows. Section 2 presents the general features of the display we propose. In section 3, we introduce some examples of the movie, and how it can provide unexpected insight into chain behavior that would not have been apparent otherwise. Finally, in section 4, we give some conclusions.

2 The General Nature of the Proposal

As described above, trace plots and other methods for evaluating the behavior of an MCMC run tend to look at each parameter marginally, whereas it is the joint behavior of all the parameters that is of true interest. The difficulties in visualizing and assessing a chain are exacerbated in more recent advances such as reversible jump MCMC, since the meaning of the various parameters can (and often will) change with the dimension of the model being fit to the data. Marginal or univariate diagnostics, such as have been suggested for ordinary MCMC, will tend to be misleading because of this. Consider for example a mixture model, with either one component or two. Suppose interest centers on the posterior distribution of the mean of each component, namely, μ_1 in the first case, and (μ_1, μ_2) in the second. The posterior distribution of μ_1 mixes the results of two models with different dimensions, for which the parameter has different interpretations. A naive approach would be to display the results marginally, conditional on the number of components in the mixture, and there is some information in such plots. However, they don't permit evaluation of global qualities of the chain, such as mixing, nor do they give much guidance on the changes in the model fit as the chain traverses the multidimensional and highly complex parameter space.

To get around these various obstacles, we propose a simple method of display, which is a multivariate extension of the trace plot. The status of some (but not all) MCMC runs at a given instant in the chain can be summarized by a plot. For example, a univariate piecewise regression with an uncertain number of breaks can be represented by a graph that displays the data points and the fit at any given stage of the chain. Less informatively, a regression with several independent variables against a univariate dependent variable might plot the data against the fitted values, where the fit might involve several regimes of uncertain number and boundaries.

For any such summary, the stringing together of the plots, displaying development of the chain over time, makes a movie. This paper explores the utility of such movies. The key to having a good movie is having an efficient and informative summary of the model, a "clear portrayal of complexity" (Tufte, 1983, pg. 191). Some thought therefore needs to be dedicated to finding an appropriate graphical representation, which displays as many of the parameters as possible; in particular, the movie should enable the viewer to see changes that unfold in the form of the model as the dimension switches, when this is a relevant consideration. In this sense, while the tool could be useful for any MCMC problem where the output at any given iteration may be summarized in a single graph, we envision it to be especially helpful for examining the results of a reversible jump chain. The movie can be enhanced to include further dimensions of the problem by the use of color in the plots (as in one implementation described below, where models of different dimensions were plotted in different colors) or perhaps sound.

This approach is intrinsically a multivariate one, as it makes possible the examination of the posterior distributions of several parameters at once, including the dimension of the model being fit. As such, it avoids some of the problems of marginal or univariate displays. It is easy to interpret, and, as described below, interesting features of the MCMC simulation and the posterior distribution of the parameters are not hard to extract from watching the movie as it runs. We should note, however, that the method is not completely general. Rather, it is intended for use in problems where a plot of the model and its parameters, perhaps with the data, is meaningful. For example, piecewise linear models with multiple changepoints may be plotted in this way to great effect, as can spline models with a variable (or fixed) number of knots. Application of the graphical tool requires the statistician to be familiar enough with the data in question, to be able to devise a suitable presentation of the results of the MCMC run.

3 Examples

In this section, we consider one example in some depth, and several others in less detail. The first example involves the use of reversible jump MCMC for a piecewise linear model, where the number of changepoints is unknown. The application is in archaeology, for the analysis of corbelled domes. An analysis of the data set we discuss here also appears in Fan and Brooks (2000). The technique of corbelling refers to a particular manner in which stones are placed to span or roof spaces in order to create a dome–like structure, and was widely used in antiquity, before the true dome was known. Figure 1 shows an idealized version of a corbelled dome. Examples of corbelled structures have been found

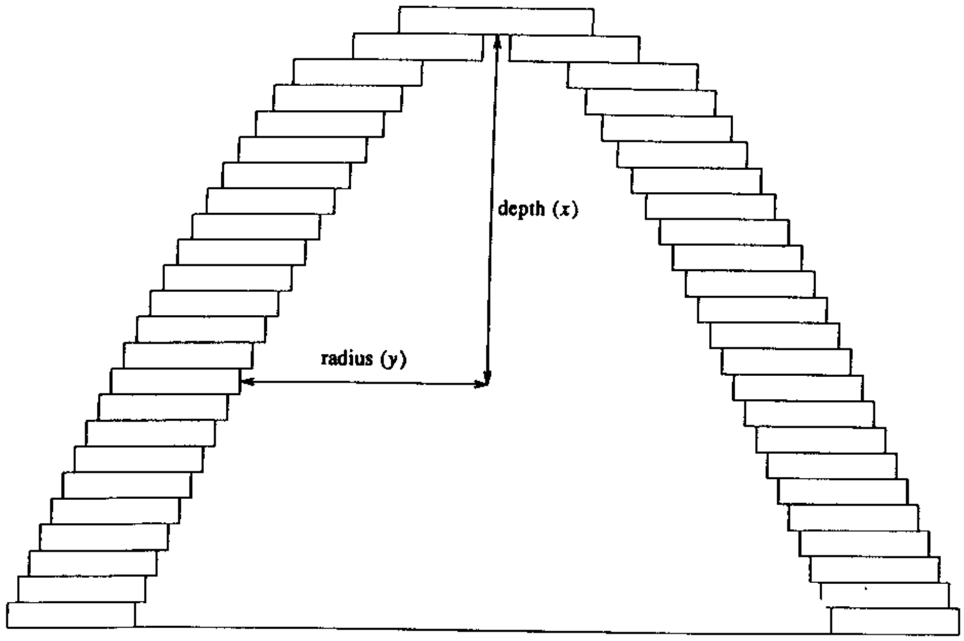

Figure 1: Shape of the inner face of an idealized corbelled structure. Reprinted from Buck, Litton and Stephens (1993) with permission from the Royal Statistical Society.

in many ancient societies, ranging from Western Europe to Egypt. Given this spatial dispersion of location, and the fact that some of the corbelled domes have survived for three millennia or more, understanding the mechanics of corbelling has been of interest to archaeologists.

Some of the finest examples of corbelling are the Mycenaean and late Minoan *tholos* tombs of Greece and Crete, respectively. The first in–depth study of the Mycenaean tholos was in Cavanagh and Laxton (1981), who analyzed the Treasury of Atreus (perhaps the most famous and well–preserved corbelled tomb from ancient times), as well as other tholoi in Mycenae, from the perspective of structural mechanics. These authors aimed to uncover what was necessary for a corbelled tomb to stand, and remain standing over the centuries. Cavanagh and Laxton proposed a simple model to determine the curve of the internal surface of the dome, namely,

$$y = \alpha x^{\beta},$$

where α and β are constants to be determined for each tomb, x is the depth as measured down from the highest point in the dome (more on this below), and y is the radius as measured at each measured depth (see Figure 1). Putting this on the log scale, we have that

$$\log y = \beta \log x + \log \alpha.$$

By comparing the values of α and β, it is possible for archaeologists to form theories about the spread of knowledge across ancient peoples pertaining to the conditions needed for stable corbelled domes. This model was deterministic and was found to fit the tholoi at Mycenae quite well; furthermore, the values of β for the five tombs examined in this paper were all similar (around 0.7), while the values of α ranged from 1.2 to 2.7.

In a later paper (Cavanagh, Laxton and Litton, 1985), two extensions were added – the possibility of a change of slope and intercept was considered, making this now a changepoint problem and introducing a stochastic element; and allowing for an *offset* in the measurements of the depth, resulting from the exact apex of the dome being unknown (this could have a number of causes, such as the top having collapsed, or the tomb not having been completed with smaller stones, with the builders instead laying down a large capstone). The new model is therefore

$$\begin{aligned}
\log y_i &= \log \alpha_1 + \beta_1 \log(x_i + \delta) + \epsilon_i && x_i \leq \gamma \\
\log y_i &= \log \alpha_2 + \beta_2 \log(x_i + \delta) + \epsilon_i && x_i \geq \gamma
\end{aligned}$$

where x_i is the depth at the ith measured point, y_i is the corresponding radius, γ is the location of the changepoint, δ is the offset, the difference between the theoretically true top of the dome and the spot where measurements could begin in reality, and the ϵ_i terms are independent, $N(0, \sigma^2)$. This model was fit on the five tombs from Cavanagh and Laxton (1981), as well as Minoan tholoi and corbelled tombs called *nuraghi* from Sardinia. Fitting the expanded model highlighted regional differences – whether or not there was evidence for a changepoint, the location of the changepoint along the dome, and the estimated slopes, intercepts and offsets varied for the nuraghi compared to the tholoi, and even within these broader divisions.

A Bayesian analysis of this model was performed by Buck, Litton and Stephens (1993) on the tholos tomb at Stylos, Crete. These authors stipulated a single changepoint, but there is no *a priori* reason to think that one change and only one change occurs. While the number of changepoints is likely to be small, one can conceive of multiple changes taking place as the building of the dome progressed.

We investigate this latter option on the data from the tomb at Stylos, an example of a late Minoan (around 1500 to 1000 B.C.E.) tholos tomb. The model we consider is of the form

$$
\begin{aligned}
\log y_i &= \log \alpha_1 + \beta_1 \log(x_i + \delta_1) + \epsilon_i && x_i \leq \gamma_1 \\
\log y_i &= \log \alpha_2 + \beta_2 \log(x_i + \delta_2) + \epsilon_i && \gamma_1 \leq x_i \leq \gamma_2 \\
&\vdots \\
\log y_i &= \log \alpha_{k+1} + \beta_{k+1} \log(x_i + \delta_{k+1}) + \epsilon_i && x_i \geq \gamma_k
\end{aligned}
$$

allowing for up to k changepoints. Using conventional methods, with k being itself a parameter of the problem, would be extremely difficult, so we opt for a Bayesian analysis. Fixed–parameter size MCMC techniques are also inapplicable, hence we adopt the reversible jump methodology, allowing as many as three changepoints. Since there are only 24 measurements in this data set, we did not want to allow more than four separate segments to be fit. Furthermore, questions of interpretation could arise with more changepoints, if they did not correspond to structural features in the tomb. We follow the algorithm given in Fan and Brooks (2000); see there for details of implementation.

In thinking about monitoring, displaying or assessing the results of our reversible jump MCMC chain, we encounter the difficulty that the meanings of some of the parameters differ for models of differing dimensions, as discussed in Section 2. For instance, the interpretation of the slope parameters for each line is different, depending on how many changepoints are in the model. Indeed, even the interpretation of the changepoints themselves could vary – γ_2 doesn't have the same meaning in a model with two changepoints as it does in a model with three, and for models with fewer than two changepoints, this parameter isn't even there to interpret. Comparing across models of different dimensions, in order to monitor convergence or the mixing of the chain, is thus difficult. In some sense this is a labeling issue, as mentioned in Richardson and Green (1997), but it also has very real implications in designing tools for looking at the data. The movie is a way of investigating some of these questions.

A crucial step is to find a suitable graphical display of the underlying model connecting the parameters. In the tomb example, this is not hard to do – just plot the data points, $\log x_i$ on the x axis and $\log y_i$ on the y axis, connecting these with the fitted lines defined by the changepoints. Models of different dimensions, represented by varying numbers of changepoints, are plotted with lines of different colors, and the changepoints are also plotted in a contrasting color to make them stand out. For each iteration in the chain, the data and the fitted model are plotted.

This movie makes it possible to monitor the mixing and convergence of the chain along various parameters of interest – by observing changes in color, for instance, one can see that the chain spends most of its time visiting models with three changepoints. While this observation can be made by looking at a marginal trace plot for the model dimension, the movie allows us to see in addition how the locations of the changepoints jump from model to model, how changes in the offset parameters δ are reflected in the fits, and how the slopes and intercepts of the individual lines vary to account for all of these simultaneous moves in the values of the other parameters. We can note for example that models with no changepoints result in poorer fits overall. Models with one or two changepoints that are generated shortly after a zero changepoint model tend to also have poor fits; on the other hand, one or two changepoint models that arise from higher dimensional models have better fits. In addition, as might be hoped, the model tends to add changepoints where they are most needed. If a changepoint is added very close to an already–existing changepoint, the fit is somewhat worse, apparently due to adding in a superfluous segment. Our movie thus offers a dynamic way of viewing the results of the MCMC run. Furthermore, all of the parameters of interest, with the exception of the error variance, σ^2, are presented in the movie (the offset parameters only implicitly), although we are able to get at least a qualitative impression of this parameter as well. We also created a version of the movie that includes σ^2, by plotting one standard deviation bars around the data. While there is information in this movie, we feel that it is still not a satisfactory way of including this parameter, as the result is more cluttered and harder to follow.

Using the movie technique it is very easy to make changes in the burn–in period (simply start the movie from a later or earlier frame and watch the differences), and to thin the chain (by only showing every rth frame, for any value of r we might choose). Therefore multiple aspects of the chain can be monitored at once. In an early implementation, for example, we noticed that there was a fair amount of "clumping" even in a multivariate sense, that is, there were portions of the chain characterized by unchanging dimension and δ, with little movement in the existing changepoints. Plotting only every 10th frame alleviated this problem and revealed the important features of the fitted models.

The movie also displays other characteristics that might have been difficult to spot graphing the chain parameter by parameter. For instance, we are able to monitor where changepoints tend to appear and disappear, both relative to those currently in the model, and relative to location on the cross–section of the tomb. The chain concentrates most of its effort on trying to fit the data well near the bottom, by placing changepoints in the lower half of the data more than in the upper half. A possible explanation for this is that the builders of the tomb had to make adjustments near the bottom to accommodate features of the ground in which the structure was placed – many Mycenaean tombs were built inside a hole dug into the earth, or inside mounds that helped provide stability, and this would need to be taken into account in the early stages of construction. Near the top, fewer adjustments would be necessary.

The flexibility of the movie also frees us to explore some of the more exotic aspects of Markov Chain Monte Carlo simulation. For example, one might perform sensitivity analyses under the sampling–resampling perspective of Smith and Gelfand (1992), where the importance weight for going from one prior to another could be represented by the length of the frame (equivalently, and perhaps easier to program, the number of consecutive times the frame is shown). Another use for this idea would be to sample under one prior that produces a chain that converges quickly and mixes well, but to interpret the results under another prior, that more realistically summarizes the user's beliefs about the state of nature, as was done in DiMatteo and Kadane (2001).

The prior we used for the movie is the prior we are interested in for the analysis of the data. But had we wished to emphasize models with only one or two changepoints, for example, we could have run the chain with a prior that artificially increases the probability on that part of the parameter space. The chain would then be allowed to sample there more heavily, but the parameter points with one or two changepoints could be downweighted using sampling–resampling, which amounts to importance sampling in this context. Expressing the reweighting proportionally to the time spent displaying a particular set of parameters leads to movies that are equivalent, up to sampling fluctuations, to those run on the prior of interest.

On a more pragmatic note, the movie can also be used for debugging. An early run of the movie revealed that the piecewise linear segments were not matching up at the changepoints, which is required by the model (and is obviously of structural importance, as the dome wouldn't stand if there were discontinuities). Simply looking at posterior distributions of the individual parameters would not have uncovered this break in the model specification, nor would any of the usual diagnostics applied to the output of the chain. Seltman (1999) also used a movie for debugging purposes, in his case to diagnose poor mixing of a chain, and claimed that without this dynamic method, it would have been difficult, if not impossible, to understand what was causing the problem, and to work out ways of addressing it.

Another movie we have seen includes one for spline fitting, where the number and location of the knots are unknown (Wallstrom, Personal Communication). In this example, the data are plotted along with the spline fit, and all of the knots. The number of knots at each iteration can be plotted at the top of the graph, which provides a running summary of the evolution of this parameter. In this manner it can be seen how quickly the number of knots changes (an indication of mixing), how the locations of the knots are affected by the number of knots, and so on. If a goodness–of–fit measure is also included, the user can monitor how increasing or decreasing the number of knots will affect the posterior variance of the model.

A general sketch of the method is as follows:

1. Choose or find a meaningful graphical summary of the model at a given instant in the running of the chain. Some creativity might be needed here! For particular situations, such as regression and related models, the choice is relatively straightforward. Other models might require more thought. In any case, the summary that is chosen should incorporate as many of the parameters as is feasible, and should provide the viewer (user) with a snapshot of the model at any given stage of the chain.

2. Set a burn–in period of B iterations, and a thinning factor of r. These may be determined theoretically (see for example, Raftery and Lewis, 1992), from inspection of trace plots, or from an initial viewing of the movie based on all iterations of a pilot simulation study.

3. For every rth iteration after B (r might be 1), plot the graphical summary from step (1), using the parameter values for that iteration.

As should be clear, programming the movie is not at all difficult – a few lines of code in Splus or any other software package that can produce graphs and loop through them, are all that is required. An important design question that we have encountered is speed; how quickly the program cycles through the frames of the movie needs to be controlled, since if it goes too quickly, it is impossible to discern interesting features and patterns, whereas if it moves too slowly, it can be hard to form impressions about the status of the chain. Mechanisms for varying the speed of the movie will differ according to the platform (Java as opposed to graphics window in Splus, for instance). Whether or not to redraw the axes at each iteration was an unexpected choice that we had to face. Redrawing the axes for each frame of the movie may produce a distracting flicker effect, which can make it harder to concentrate on the unfolding of the simulation. The principles of good graphics, put forth by Tukey (1977) and Tufte (1983), should guide the design not only of the individual frames of the movie, but also of the display taken as a dynamic whole.

Simple regression and related methods, for instance, piecewise linear (multiple change-point) and spline (fixed or variable number of knots) models, can all be handled in a similar way – plot the data, with the fitted model drawn over them. This particular approach is suitable only for relatively simple models; multiple regression, for example, involves high–dimensional spaces which cannot be plotted as a single snapshot of the model. Other summary graphics are required for this case. One possibility was already mentioned, namely, to plot the data and the fitted values together; another is to plot the estimated coefficients at each iteration on a barplot (with positive coefficients pointing up above the zero horizon, negative coefficients pointing down), allowing the viewer to monitor how they change as the Markov chain progresses, whether there are coefficients that aren't being sampled efficiently, and so on. The variance estimate $\hat{\sigma}^2$ could be given explicitly at the bottom of the plot, or incorporated into the barplot. This could also be used as a tool for variable selection, as the user can observe which coefficients, if any, are consistently estimated to be near zero.

Moving away from regression, we consider two other commonly–used models, mixture models and time series. Both of these lend themselves to graphical representation, again keeping in mind the general principle of plotting the data with the fitted model. For mixture models, a possible snapshot is given in Figure 2. The data in this figure are prices (in pounds Sterling) of the thirty–one different wooden toys on sale at a craft shop in Suffolk in April 1991, and are taken from Hand *et al.* (1994). In this display, the data are presented in a histogram, and the fitted mixture model is drawn over the histogram. This picture allows the user to see how the fit changes as the mean, variance, and weight of each component moves along with the chain. If the number of components is also a parameter of the model, that is, the number of components is not fixed but rather is allowed to vary, it is possible further to see how the fit is affected by the number of components in the mixture. The plot can be enhanced by adding a counter to the plot, that keeps track of the number of components fit. Other embellishments can be included, however the ease with which the movie as a whole is grasped should be kept in mind.

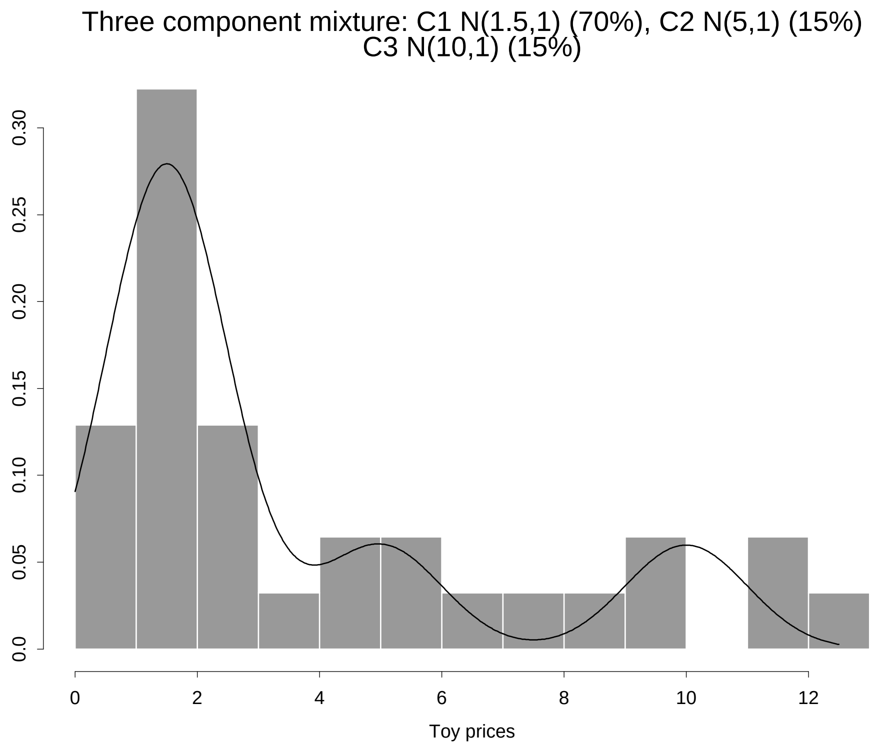

Figure 2: Snapshot of movie for toy prices data. A three component normal mixture is fit to the data, which appear in the histogram. The first component, with weight 70%, has mean 1.5; the second, with weight 15%, has mean 5; the third component, with weight 15%, has mean 10. All mixture components have variance 1.

For the visualization of time series models, consider the luteinizing hormone data from Diggle (1990), also analyzed as an example in Venables and Ripley (1997). The data are plotted in Figure 3, along with a fitted autoregressive model of order 1. Successive snapshots in this case would show how the fit of the model changes as the value of the autoregressive parameter shifts over samples in the MCMC run. Adjustments to the basic plot in this case include the use of color to reflect the fitting of autoregressive models of different orders, different line types to represent models that are solely autoregressive, solely moving average, or ARMA (assuming that the chain being run is flexible enough to encompass all of these possibilities) and the addition of a legend with the estimated variance. Again, other features can be added at the user's discretion.

4 Conclusion

We propose a simple graphical display for MCMC output – a statistical movie. A main feature of the proposal is that it can be used for problems where the data and the model can be effectively summarized in a single plot. This display has the advantages of presenting the model and the data together in a manner that is easily interpreted, and of incorporating the posterior distributions of several parameters at once. Its dynamic nature allows the researcher to follow the chain as it traverses the joint posterior parameter space, to diagnose problems such as multivariate clumping and poor mixing, and to discover aspects of the fitted models that might not otherwise be readily apparent. It also has the potential to offer convenient sensitivity analysis for complicated models, through

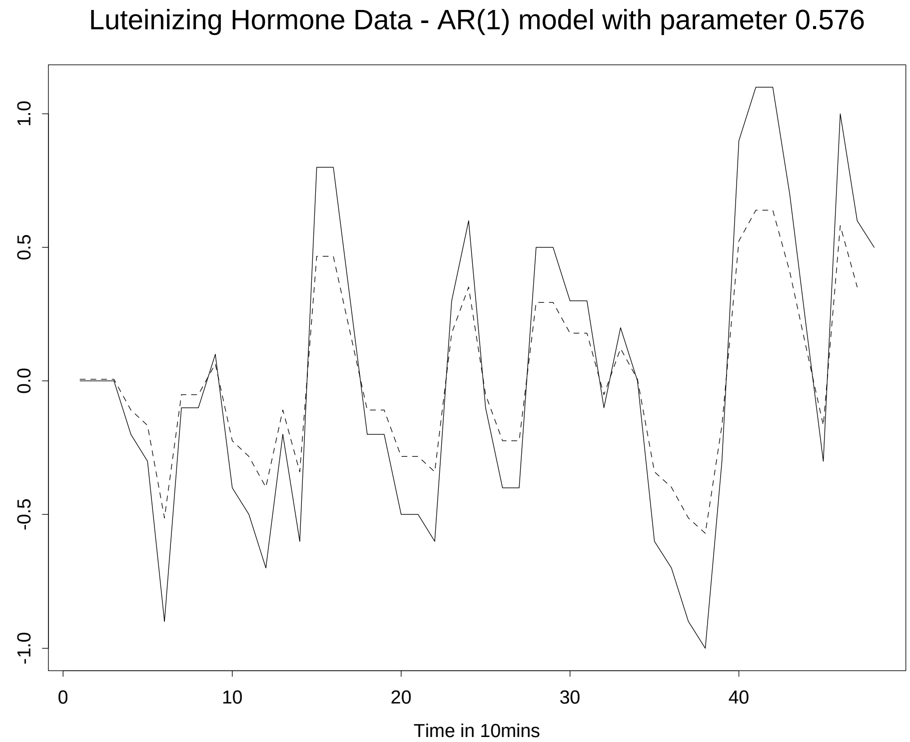

Figure 3: Snapshot of movie for hormone data. The data are drawn in the solid line; an autoregressive model of order 1 is fit to the data and drawn with the dashed line. The parameter of the AR(1) model is 0.576.

the use of sampling–resampling and appropriate reweighting of the movie frames. Use of the statistical movie requires the researcher to find a suitable way of graphically summarizing the data and the parameters, which will be problem specific. We have given suggestions for some common models, although others are also possible. While we anticipate that for many applications, an appropriate summary will suggest itself naturally, it is conceivable that for complicated or very high–dimensional models, there will be no such convenient way of capturing multiple parameters at once in a single plot. Alternatively, in the same situation several summaries might be available, in which case perhaps a split–screen movie would be informative.

Acknowledgments: The authors would like to thank Howard Seltman for pointing out his use of the movie idea for diagnosing mixing problems, Garrick Wallstrom for implementing the movie and sharing the results with us, and Caitlin Buck for helpful discussions about corbelling and statistics in archaeology. Several anonymous reviewers read an earlier version of this paper; their comments are appreciated. We also thank the Royal Statistical Society for permission to reprint Figure 1. This work was partially supported by NSF grants (Kadane) DMS–9802401 and DMS–9705034 (Lazar).

REFERENCES

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

Brooks, S.P. and Giudici, P. (2000) Markov Chain Monte Carlo convergence assessment via two-way analysis of variance. *Journal of Computational and Graphical Statistics*, **9**, 266–285.

Buck, C.E., Litton, C.D. and Stephens, D.A. (1993) Detecting a change in the shape of a prehistoric corbelled tomb. *The Statistician*, **42**, 483–490.

Cavanagh, W.G. and Laxton, R.R. (1981) The structural mechanics of the Mycenaean tholos tomb. *Annual of the British School at Athens*, **76**, 109–140.

Cavanagh, W.G., Laxton, R.R. and Litton, C.D. (1985) An application of change–point analysis to the shape of prehistoric corbelled domes; I. The maximum likelihood method. *PACT*, **11 – III.4**, 191–203.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Chib, S. and Jeliazkov, I. (2000) Marginal likelihood from the Metropolis–Hastings algorithm. Washington University Technical Report.

Cowles, M.K. and Carlin, B.P. (1996) Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Diggle, P.J. (1990) *Time Series: A Biostatistical Introduction.* Oxford, U.K. Oxford University Press.

DiMatteo, I. and Kadane, J.B. (2001) Vote tampering in a district justice election in Beaver County, PA. *Journal of the American Statistical Association*, **96**, 510–518.

Fan, Y. and Brooks, S.P. (2000) Bayesian modelling of prehistoric corbelled tombs. *The Statistician*, **49**, 339–354.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Geweke, J. (1992) Evaluating the accuracy of sampling–based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, Bernardo, J.M., Berger, J., Dawid, A.P. and Smith, A.F.M. (eds), Oxford, U.K., Oxford University Press, 169–193.

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1994) *A Handbook of Small Data Sets.* London, Chapman and Hall.

Raftery, A. and Lewis, S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, Bernardo, J.M., Berger, J., Dawid, A.P. and Smith, A.F.M. (eds), Oxford, U.K., Oxford University Press, 765–776.

Richardson, S. and Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

Seltman, H. (1999) Hidden stochastic models for biological rhythm data. Unpublished PhD thesis, Department of Statistics, Carnegie Mellon University.

Smith, A.F.M. and Gelfand, A.E. (1992) Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician*, **46**, 84–88.

Tufte, E.R. (1983) *The Visual Display of Quantitative Information* Cheshire, Graphics Press.

Tukey, J.W. (1977) *Exploratory Data Analysis* Reading, Addison–Wesley.

Venables, W.N. and Ripley, B.D. (1997) *Modern Applied Statistics with S–Plus* (Second ed.) New York, Springer.

Yu, B. and Mykland, P.A. (1998) Looking at Markov samplers through CUSUM path plots: A simple diagnostic idea. *Statistics and Computing*, **8**, 275–286.

LEGENDS FOR FIGURES

Figure 1 Shape of the inner face of an idealized corbelled structure. Reprinted from Buck, Litton and Stephens (1993) with permission from the Royal Statistical Society.

Figure 2 Snapshot of movie for toy prices data. A three component normal mixture is fit to the data, which appear in the histogram. The first component, with weight 70%, has mean 1.5; the second, with weight 15%, has mean 5; the third component, with weight 15%, has mean 10. All mixture components have variance 1.

Figure 3 Snapshot of movie for hormone data. The data are drawn in the solid line; an autoregressive model of order 1 is fit to the data and drawn with the dashed line. The parameter of the AR(1) model is 0.576.