

Discussion on paper by Martin Weinberg regarding “Bayesian Model selection and Parameter Estimation.”

Philip C. Gregory

Abstract Bayesian model selection and parameter estimation is attracting a lot of interest in the astronomical community because of its power and logical consistency. Markov chain Monte Carlo provides the computational power for Bayesian parameter estimation problems in large parameter spaces but needs to be supported with other numerical techniques for efficient exploration of multi-modal probability distributions. Bayesian model selection is easy in concept but remains a difficult challenge for large parameter spaces. My comments are based on lessons learned from developing a controlled statistical fusion approach to some of these issues.

Key words: Bayesian parameter estimation, Bayesian model selection, Markov chain Monte Carlo, fusion MCMC, controlled statistical fusion

1 Introduction

Martin Weinberg reports on using the UMass Bayesian Inference Engine (BIE) package for model selection and parameter estimation in extragalactic astronomy. I have independently developed a Bayesian approach to accomplish similar goals with particular emphasis in the arena of exoplanets. This conference provides an opportunity to exchange lessons learned. Not surprisingly, because of the large number of model parameters involved, both groups employ a Markov chain Monte Carlo (MCMC) integration engine. The BIE philosophy is that there is no single best MCMC algorithm and develop a variety of MCMC algorithms augmented by different tools like parallel tempering, simulated annealing and differential evolution depending on the complexity of the problem. My approach has been to attempt to fuse together the advantages of all of the above tools together with a genetic crossover operation in a single MCMC algorithm to facilitate the detection of a global minimum in χ^2 .

My latest algorithm is called fusion MCMC [12]. This fusion has only been possible through the development of a unique adaptive control system to automate the choice of an efficient set of MCMC proposal distributions even if the parameters are highly correlated. The control system also supervises the operation of the different components. Figure 1 shows two schematics on the operation of an 8 parallel chain

Philip C. Gregory
Physics and Astronomy, University of British Columbia, 6224 Agricultural Rd, Vancouver, B. C.
V6T 1Z1 e-mail: gregory@phas.ubc.ca

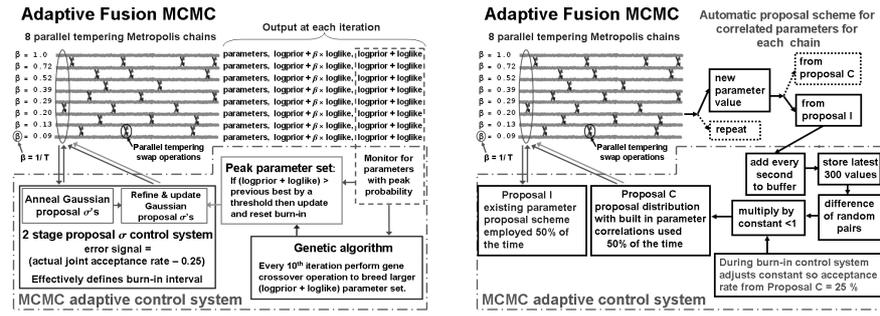


Fig. 1 Two schematics on the operation of the adaptive fusion MCMC algorithm. The right panel illustrates the automatic proposal scheme for handling correlated parameters.

fusion MCMC and the control system. In applications to real precision radial velocity data the algorithm has proved highly effective [6, 7, 9, 10, 12]. The *Mathematica* based parallel code is run on a 8 core PC and requires 10 hours for a 6 planet model with 37 parameters and one million iterations. The execution time scales with the number of planets.

2 Some Useful Lessons

2.1 Highly correlated parameters

For some models the data is such that the resulting estimates of the model parameters are highly correlated and the MCMC exploration of the parameter space can be very inefficient. One solution to this problem is Differential Evolution Markov Chain (DE-MC) [2]. DE-MC is a population MCMC algorithm, in which multiple chains are run in parallel, typically from 15 to 40, although Weiner's experience suggests that 64 chain would be the bare minimum. DE-MC solves an important problem in MCMC, namely that of choosing an appropriate scale and orientation for the jumping distribution.

For the fusion MCMC algorithm, I developed and tested a new method [11], in the spirit of DE, that automatically achieves efficient MCMC sampling in highly correlated parameter spaces without the need for additional chains. The block in the lower left panel of Fig. 1 automates the selection of efficient proposal distributions when working with model parameters that are independent or transformed to new independent parameters. New parameter values are jointly proposed based on independent Gaussian proposal distributions ('I' scheme), one for each parameter. Initially, only this 'I' proposal system is used and it is clear that if there are strong correlations between any parameters the σ values of the independent Gaussian proposals will need to be very small for any proposal to be accepted and consequently convergence will be very slow. However, the accepted 'I' proposals will generally cluster along the correlation path. In the optional third stage of the control system (see right panel of Fig. 1) every second accepted 'I' proposal is appended to a corre-

lated sample buffer. There is a separate buffer for each parallel tempering level. Only the 300 most recent additions to the buffer are retained. A ‘C’ proposal is generated from the difference between a pair of randomly selected samples drawn from the correlated sample buffer for that tempering level, after multiplication by a constant. The value of this constant (for each tempering level) is computed automatically [11] by another control system module which ensures that the ‘C’ proposal acceptance rate is close to 25%. With very little computational overhead, the ‘C’ proposals provide the scale and direction for efficient jumps in a correlated parameter space.

The final proposal distribution is a random selection of ‘I’ and ‘C’ proposals such that each is employed 50% of the time. The combination ensures that the whole parameter space can be reached and that the FMCMC chain is aperiodic. The parallel tempering feature operates as before to avoid becoming trapped in a local probability maximum.

Because the ‘C’ proposals reflect the parameter correlations, large jumps are possible allowing for much more efficient movement in parameter space than can be achieved by the ‘I’ proposals alone. Once the first two stages of the control system have been turned off, the third stage continues until a minimum of an additional 300 accepted ‘I’ proposals have been added to the buffer and the ‘C’ proposal acceptance rate is within the range ≥ 0.22 and ≤ 0.28 . At this point further additions to the buffer are terminated and this sets a lower bound on the burn-in period.

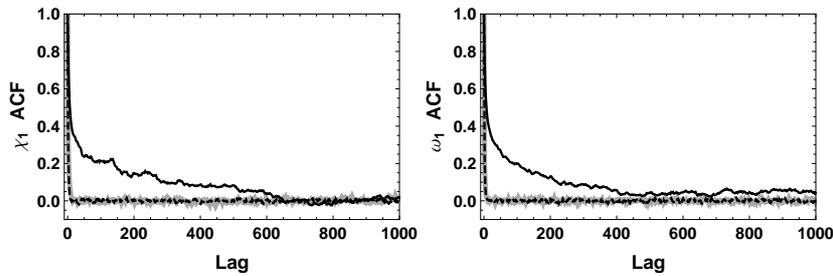


Fig. 2 The two panels show the MCMC autocorrelation functions for two highly correlated parameters χ and ω . The solid black trace corresponds to a search in χ and ω using only ‘I’ proposals. The light gray trace corresponds to a search in χ and ω with ‘C’ proposals turned on. The dashed trace corresponds to a search in the transformed orthogonal coordinates $\psi = 2\pi\chi + \omega$ and $\phi = 2\pi\chi - \omega$ using only ‘I’ proposals.

Fig. 2 shows the autocorrelation functions of post burn-in MCMC samples for two highly correlated parameters χ and ω . The solid black trace corresponds to a search in χ and ω using only ‘I’ proposals. The light gray trace corresponds to a search in χ and ω with ‘C’ proposals turned on. The dashed trace corresponds to a search in the transformed orthogonal coordinates $\psi = 2\pi\chi + \omega$ and $\phi = 2\pi\chi - \omega$ using only ‘I’ proposals. It is clear that a search in χ and ω with ‘C’ proposals turned on achieves the same excellent results as a search in the transformed orthogonal coordinates ψ and ϕ using only ‘I’ proposals.

2.2 Noise model

Based on their results, Weinberg concludes that a data-model comparison without an accurate error model is likely to be erroneous. I have found it very useful to incorporate an extra noise parameter, s , that can allow for any additional noise beyond the known measurement uncertainties¹. We assume the noise variance is finite and adopt a Gaussian distribution with a variance s^2 . Thus, the combination of the known errors and extra noise has a Gaussian distribution with variance $= \sigma_i^2 + s^2$, where σ_i is the standard deviation of the known noise for i^{th} data point. In general, nature is more complicated than our model and known noise terms. Marginalizing s has the desirable effect of treating anything in the data that can't be explained by the model and known measurement errors as noise, leading to more conservative estimates of the parameters. See Sections 9.2.3 and 9.2.4 of [1] for a tutorial demonstration of this point. If there is no extra noise then the posterior probability distribution for s will peak at $s = 0$. Incorporating an extra noise parameter also results in an auto-

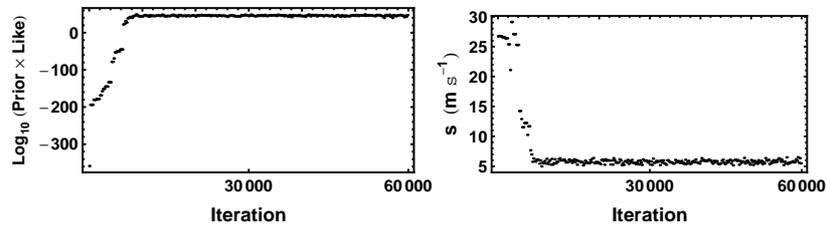


Fig. 3 The left panel is a plot of the $\text{Log}_{10}[\text{Prior} \times \text{Likelihood}]$ versus MCMC iteration. The right panel is a similar plot for the extra noise term s . Initially s is inflated and then rapidly decays to a much lower level as the best fit parameter values are approached.

matic annealing operation whenever the Markov chain is started from a location in parameter space that is far from the best fit values. When the χ^2 of the fit is very large, the Bayesian Markov chain automatically inflates s to include anything in the data that cannot be accounted for by the model with the current set of parameters and the known measurement errors. This results in a smoothing out of the detailed structure in the χ^2 surface and, as pointed out by [4], allows the Markov chain to explore the large scale structure in parameter space more quickly. The chain begins to decrease the value of the extra noise as it settles in near the best-fit parameters. An example of this is shown in Fig. 3. This is similar to simulated annealing, but does not require choosing a cooling scheme.

¹ In the absence of detailed knowledge of the sampling distribution for the extra noise, we pick an independent Gaussian model because for any given finite noise variance it is the distribution with the largest uncertainty as measured by the entropy, i.e., the maximum entropy distribution [13, 1].

2.3 Model selection

One of the great strengths of Bayesian analysis is the built-in Occam's razor. More complicated models contain larger numbers of parameters and thus incur a larger Occam penalty, which is automatically incorporated in a Bayesian model selection analysis in a quantitative fashion (see [1] for example, p. 45). Bayesian model selection relies on the ratio of marginal likelihoods where the marginal likelihood is the weighted average of the conditional likelihood, weighted by the prior probability distribution of the model parameters and any unknown additional noise parameter. At the last SCMA conference Clyde et al. [3] reviewed the state of techniques for model selection from a statistics perspective and Ford and Gregory [5] evaluated the performance of a variety of marginal likelihood estimators in the exoplanet context. The bottom line is that Bayesian model selection is easy in concept but becomes progressively more difficult to compute as the number of model parameters increase. Here we compare recent results obtained from two different methods: (1) nested restrictive Monte Carlo (NRMC), and (2) the ratio estimator (RE).

Nested restrictive Monte Carlo (NRMC) is a recent improvement [10, 12] on the RMC method. In RMC [5], the volume of parameter space sampled is restricted to a region delineated by the outer borders (e.g., 99% credible region) of the MCMC marginal parameter distributions for the dominant mode. In principle, the contribution from a secondary mode can be computed in a like fashion. In NRMC integration, multiple boundaries are constructed based on credible regions ranging from 30% to $\geq 99\%$, as needed. The contribution to the total integral from each nested interval is computed. For example, for the interval between the 30% and 60% credible regions, we generate random parameter samples within the 60% region and reject any sample that falls within the 30% region. Using the remaining samples we can compute the contribution to the NRMC integral from that interval.

The left panel of Figure 4 shows the contributions from the individual intervals for 5 repeats of the NRMC evaluation for a 3 planet model fit to the Gliese 581 [12] exoplanet system. The right panel shows the summation of the individual contributions versus the volume of the credible region. The credible region listed as 9995% is defined as follows. Let X_{U99} and X_{L99} correspond to the upper and lower boundaries of the 99% credible region, respectively, for any of the parameters, with X_{U95} and X_{L95} similarly defined. Then $X_{U9995} = X_{U99} + (X_{U99} - X_{U95})$ and $X_{L9995} = X_{L99} + (X_{L99} - X_{L95})$. Similarly, $X_{U9984} = X_{U99} + (X_{U99} - X_{U84})$.

Table 1 shows a comparison of the NRMC method to a second marginal likelihood estimator called the Ratio Estimator [5] (RE), for three planet (17 parameters) and four planet (22 parameters) exoplanet models for three different stars HD 11964, 47 UMa, and Gliese 581. The RE method employed a mixture of 150 multivariate Normals [9] to approximate the MCMC samples. The latest version improves the handling of wrap around angular parameters in the calculation of the covariance matrix of each multivariate Normal. For the three planet models the NRMC and RE methods agree within 25%. In the case of HD11964, one of the 3 signals is a suspected artifact but this is of no consequence for the present comparison of marginal likelihood estimators. At sufficiently high dimensions, the NRMC method

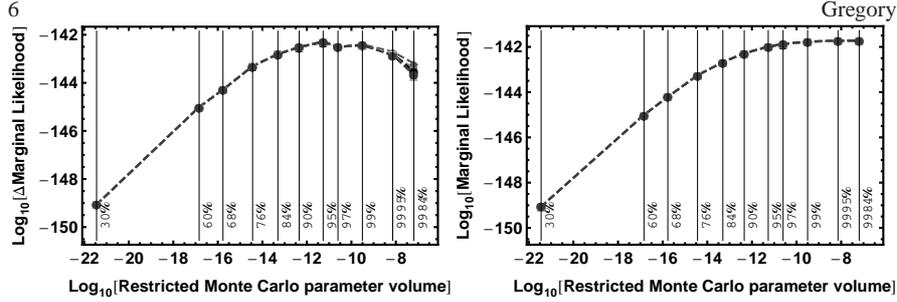


Fig. 4 Left panel shows the contribution of the individual nested intervals to the NRMC marginal likelihood for the 3 planet model (17 parameters) for 5 repeats. The right panel shows the integral of these contributions versus the parameter volume of the credible region.

is expected to underestimate the marginal likelihood and the factor by which it underestimates is expected to grow with increasing dimension. Thus NMRC estimated Bayes factor should not falsely support a more complicated model and in this sense the NRMC method is expected to fail in a conservative fashion. On the other hand, the RE method has the potential to pay too much attention to the mode as each integrand in the ratio involves the square of the posterior density and is expected to overestimate the marginal likelihood at sufficiently high dimensions. As the table indicates, by the time we reach a 4 planet model (22 parameters) one or both of these methods is failing.

Table 1 The ratio of the NRMC and RE marginal likelihoods estimates for three planet (17 parameters) and four planet (22 parameters) exoplanet models.

star	# planets	NRMC Estimator
		RE Estimator (improved version)
HD 11964	3	0.9
47 UMa	3	0.75
Gliese 581	3	1.01
Gliese 581	4	0.016

3 Acknowledgments

The author would like to thank Wolfram Research for providing a complementary license for gridMathematica.

References

1. Gregory, P. C.: Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with *Mathematica* Support, Cambridge University Press (2005)
2. Ter Braak, C. J. F.: A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces *Statistical Computing*, 16, 239–249 (2006)
3. Clyde, M. A., Berger, J. O., Bullard, F., Ford, E. B., Jeffreys, W. H., Luo, R., Paulo, R., Lored, T.: Current Challenges in Bayesian Model Choice. In ‘Statistical Challenges in Modern Astronomy IV,’ G. J. Babu and E. D. Feigelson (eds.), ASP Conf. Ser., 371, 224–240 (2007)
4. Ford, E. B.: Improving the Efficiency of Markov Chain Monte Carlo for Analyzing the Orbits of Extrasolar Planets. *ApJ*, 620, 481 (2006)
5. Ford, E. B., & Gregory, P. C.: Bayesian Model Selection and Extrasolar Planet Detection. In ‘Statistical Challenges in Modern Astronomy IV,’ G. J. Babu and E. D. Feigelson (eds.), ASP Conf. Ser., 371, 189–204 (2007)
6. Gregory, P. C.: A Bayesian Analysis of Extrasolar Planet Data for HD 73526. *ApJ*, 631, 1198–1214 (2005)
7. Gregory, P. C.: A Bayesian Kepler Periodogram Detects a Second Planet in HD 208487. *MNRAS*, 374, 1321–1333 (2007)
8. Gregory, P. C., in ‘Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 27th International Workshop’, Saratoga Springs, eds. K. H. Knuth, A. Caticha, J. L. Center, A. Giffin, C. C. Rodriguez, AIP Conference Proceedings, 954, 307 (2007)
9. Gregory, P. C.: A Bayesian Periodogram Finds Evidence for Three Planets in HD 11964. *MNRAS*, 381, 1607–1619 (2007)
10. Gregory, P. C., and Fischer, D. A.: A Bayesian Periodogram Finds Evidence for Three Planets in 47 Ursae Majoris. *MNRAS*, 403, 731–747, (2010)
11. Gregory, P. C.: Bayesian Exoplanet Tests of a New Method for MCMC Sampling in Highly Correlated Parameter Spaces. *MNRAS*, 410, 94–110 (2011)
12. Gregory, P. C.: Bayesian Re-analysis of the Gliese 581 Exoplanet System. *MNRAS*, in press (2011)
13. Jaynes, E. T., 1957, Stanford University Microwave Laboratory Report 421, Reprinted in ‘Maximum Entropy and Bayesian Methods in Science and Engineering’, G. J. Erickson and C. R. Smith, eds, Dordrecht: Kluwer Academic Press, p.1 (1988)
14. Jeffreys, W. H.: Discussion on “Current Challenges in Bayesian Model Choice” by Clyde et al. In ‘Statistical Challenges in Modern Astronomy IV,’ G. J. Babu and E. D. Feigelson (eds.), ASP Conf. Ser., 371, 241–244 (2007)