Computational Technology for Bayesian Inference

Thomas J. Loredo

Department of Astronomy, Cornell University

Abstract. One of the most important practical differences between the Bayesian approach to statistical inference and the more traditional frequentist approach is the nature of the sums and integrals required to implement each approach. In both approaches, the sampling distribution for the data (i.e., the likelihood function) plays a key role; but frequentist calculations integrate the sampling distribution over the sample space, whereas Bayesian calculations integrate it over hypothesis (parameter) space. The numerous advantages to working in parameter space come at a cost: the required integrals are difficult to calculate. I survey recent developments in computational technology for performing such integrals.

1. Introduction

Statistical calculations consist largely of weighted sums or integrals of probabilities. One of the key practical differences between Bayesian inference and frequentist statistics is that drastically different types of integrals appear in calculations taking these competing approaches to the same problem (Loredo 1992). Consider, for example, estimating the parameters of some model, M, given some observed data, D; denote the parameters jointly by θ. A key quantity appearing in both Bayesian and frequentist integrals is the probability for the data assuming the model to be true and presuming the parameters to be known, $p(D|\theta, M)$. Considered as a function of the data, this is called the *sampling distribution*; considered as a function of the parameters, it is called the *likelihood function*, which will will abbreviate as $\mathcal{L}(\theta)$. A basic practical difference between the methods is that frequentist calculations require integrals of this quantity over the data dimensions (sample space), whereas Bayesian calculations require integrals over the parameter space.

Basing inferences on probabilities calculated by summing or integrating over parameter space brings with it a host of advantages over trying to make inferences using probabilities calculated in sample spaces. In the brief space alotted here, significant discussion of these advantages is not possible. But two positive advantages that are of great practical utility must be mentioned. First, in the vast majority of real applications the parameter space may be divided into two parts, $\theta = (\psi, \phi)$, where interest focuses on ψ, and ϕ consists of "nuisance" parameters necessary for modeling the data but otherwise uninteresting (e.g., background intensities). In Bayesian inference one can straightforwardly eliminate the nuisance parameters while accounting for their uncertainty simply by integrating the joint distribution for (ψ, ϕ) over ϕ. There is no completely sat-

isfactory method for dealing with nuisance parameters in frequentist statistics. Second, comparison of competing models in Bayesian inference is accomplished by comparing the *average* likelihoods of the parameters for the models, rather than the *maximum* likelihoods used in frequentist statistics. The averaging of the likelihood implements an automatic "Occam's razor" that penalizes a model for the size of its parameter space. There is no counterpart to this in frequentist statistics. Interested readers can consult Loredo (1990, 1992) for more discussion of these advantages, as well as discussion of some of the "negative advantages," that is, weaknesses and problems of frequentist procedures that are avoided by calculating in hypothesis space rather than sample space.

These advantages come with a significant cost: parameter space integrals are difficult to evaluate. This paper briefly introduces modern computational technology that has made such calculations feasible in problems of realistic complexity. I presume of the reader some basic familiarity with common statistical terminology. Readers needing a basic introduction to Bayesian inference will find brief, self-contained coverage in two review articles (Loredo 1990, 1992) or in § 2 of Gregory and Loredo (1992), and a thorough pedagogical account in Devinder Sivia's introductory text (Sivia 1996). Those with access to the World Wide Web can find links to some of these works, and to other Bayesian resources of particular interest to physical scientists, at the BIPS[1] (Bayesian Inference for the Physical Sciences) web site. Evans and Swartz (1995) provide a longer survey for a statistical audience covering topics in § 3 to 5 of this article.

2. Comparing Bayesian and Frequentist Integrals

Consider the common case of parameter estimation with data, D, consisting of N statistically independent samples, d_i. Suppose the model, M, has m parameters (denoted collectively by θ), with $m \ll N$. Independence implies that the likelihood can be written as a product of N terms, $\mathcal{L}(\theta) = \prod_i p(d_i|\theta, M)$. The integrals (functionals) we would need to compute in a frequentist analysis resemble the following:

$$I[f;\theta] \equiv \int d^N D\, f(D) p(d_1|\theta, M) \cdots p(d_N|\theta, M). \tag{1}$$

For example, we can calculate the sampling distribution for some statistic $S(D)$ by taking $f(D) = \delta[S - S(D)]$, or we could evaluate the bias of an estimator $\hat{\theta}(D)$ by taking $f(D) = \hat{\theta}(D)$. A rigorous frequentist procedure usually requires that the value of such an integral not depend on θ (since we do not know what the true value of θ is to condition on). In contrast, the integrals required in a Bayesian analysis resemble the following:

$$I[g] \equiv \int d^{m'}\theta\, g(\theta) p(\theta|M) \mathcal{L}(\theta), \tag{2}$$

where $p(\theta|M)$ is a prior distribution for the parameters, and the integral of interest may be over a subset of the parameter space, so $m' \leq m$. For example,

[1] `http://astrosun.tn.cornell.edu/staff/loredo/bayes/`

$I[1]$ gives us the normalization constant for Bayes's theorem; $I[\theta]/I[1]$ gives us the posterior mean estimate for θ, and taking $g(\theta)$ to be an appropriately located "hat" function allows calculation of the probability in a credible region.

Consider first low-dimensional models, with $m \lesssim 3$. For such models, evaluation of the required integrals is usually much simpler for a Bayesian calculation because the dimension of the integral is so much smaller. Analytical integration may be possible; at worst, the integral can be calculated to any required precision with straightforward numerical quadrature. In contrast, even for the simplest problems, the frequentist calculation is not trivial because of the large dimension of the sample space; one must typically use characteristic functions or integral transforms to reduce the dimensionality. As an instructive example, the reader is invited to compare the Bayesian and frequentist treatments of estimating a Gaussian mean when the noise variance is not known a priori (the solution uses Student's t statistic; cf. Sivia 1996 and Meyer 1975 for Bayesian and frequentist derivations). The final procedures are essentially the same, but the calculations leading to them differ dramatically in complexity.

But now consider higher dimension models, with $m \gtrsim 4$. Rigorous frequentist methods usually do not exist for such models, but approximate procedures are easy to develop using Monte Carlo methods.[2] Despite the large dimension of the integrals, the independence of the d_i makes implementation of such methods almost trivial: one simply samples each d_i value independently and evaluates $f(D)$; repeating this simple procedure a large number of times and averaging the results evaluates the integral. In contrast, unless a lucky (or clever) model choice allows some or all of the dimensions to succumb to analytical integration, numerical calculation of a Bayesian integral can be extremely difficult, even though it may have orders of magnitude fewer dimensions than the frequentist one. The reason is that the likelihood is almost never a product of independent factors for each parameter—inferences for the parameters are correlated, often in complicated ways. This prevents simple Monte Carlo integration as described above. Also, the probability factors comprising the likelihood are typically simple and unimodal as a function of the d_i, whereas their product may have a complicated shape as a function of θ, perhaps with multiple modes. Straightforward numerical quadrature can solve the problem in principle, but in practice the "curse of dimensionality" makes quadrature unfeasible if there are more than 3 or 4 dimensions. The developer of a Bayesian code is in the ironic position of being able to write down the exact answer to a problem that cannot be treated exactly in the frequentist approach, but not being able to actually calculate the numerical value of the answer!

Fortunately, the last ten to fifteen years have seen remarkable developments in practical algorithms for performing Bayesian calculations. We can usefully group these algorithms into three families: asymptotic approximations; methods for moderate dimensional models; and methods for high dimensional models. We discuss these in turn. All of these methods can benefit significantly from

[2]One might legitimately complain that we are comparing approximate frequentist calculations with exact Bayesian calculations, and are thus being unfair to the Bayesian approach. Unfortunately, rigorous frequentist methods simply do not exist for many realistic problems (see, e.g., § 20.35 of Kendall and Stuart 1977); we can only compare the complexities of the calculations actually done in practice.

reparameterization of a problem to simplify the structure of the posterior; several of the references below discuss this issue.

3. Asymptotic Approximations

It is often the case that the posterior distribution has a single dominant interior mode (i.e., the mode is not on the boundary of the allowed parameter space). Call this mode $\hat{\theta}$. In the vicinity of the mode, the product of the prior and likelihood can be approximated by a multivariate Gaussian, so we have,

$$p(\theta|M)\mathcal{L}(\theta) \approx p(\hat{\theta}|M)\mathcal{L}(\hat{\theta}) \exp\left[-\frac{1}{2}(\theta-\hat{\theta})\cdot\tilde{I}\cdot(\theta-\hat{\theta})\right], \tag{3}$$

where $\tilde{I}$ is the (observed) Fisher information matrix, a matrix of second derivatives evaluated at the mode: $\tilde{I} = \partial^2 \ln\left[p(\theta|M)\mathcal{L}(\theta)\right]/\partial^2\theta$ for $\theta = \hat{\theta}$. We can find approximate Bayes factors for model comparison by using this approximation to calculation average likelihoods (normalization constants for parameter estimation):

$$\int d\theta\, p(\theta|M)\mathcal{L}(\theta) \approx p(\hat{\theta}|M)\mathcal{L}(\hat{\theta})(2\pi)^{m/2}|\tilde{I}|^{-1/2}. \tag{4}$$

We can also use the approximation to do the integrals needed to eliminate nuisance parameters. If there are uninteresting parameters, ϕ, and parameters of interest, ψ, first construct a "profile" function for ψ, found by maximizing the prior $\times$ likelihood over ϕ (for each ψ): $f(\psi) = \max_\phi p(\psi,\phi|M)\mathcal{L}(\psi,\phi)$. We can construct an approximate marginal distribution for ψ by normalizing the product of $f(\psi)$ and a factor that accounts for the volume of the ϕ space:

$$p(\theta|D,M) \propto f(\psi)|\tilde{I}(\psi)|^{-1/2}, \tag{5}$$

where $\tilde{I}(\psi)$ is the information matrix for the nuisance parameters, with ψ held fixed. This approximation improves on the profile likelihood, $\max_\phi \mathcal{L}(\psi,\phi)$, the frequentist attempt to handle nuisance parameters best-known to astronomers (see, e.g., § 15.6 of Press, et al. 1992).

The use of this kind of approximation originates with Laplace, so these approximations are called *Laplace approximations*. They can perform remarkably well in practice even for modest amounts of data, despite the fact that one might expect the underlying Gaussian approximation to be good only to order $1/\sqrt{N}$, the usual rate of asymptotic convergence to a Gaussian for frequentist approximations. The reason is that the final quantities reported in a Bayesian calculation are always *ratios* of integrals. The leading order errors in the numerator and the denominator typically cancel, so Laplace approximations are usually good asymptotically to order $1/N$ or even higher. Good entry points to the literature on this approximation can be found in Tierney and Kadane (1986), Kass, Tierney, and Kadane (1991), and O'Hagan (1994).

A particular appeal of the Laplace approximation to developers building on existing frequentist codes is that all of the ingredients, apart from the prior factor, are likely to be at hand. Many frequentist codes extremize a likelihood

or log-likelihood (e.g., χ^2), and provide an approximate covariance matrix that is usually just the inverse of the information matrix required for the Laplace approximation. The Laplace approximation thus provides a quick entry into Bayesian computation.

Finally, similar reasoning underlies a somewhat cruder approximation that provides a Bayesian counterpart to the familiar frequentist procedure of comparing nested models using likelihood ratios (e.g., differences in minimum χ^2) and counting degrees of freedom. Asymptotically, the Type I error ("false alarm") probability associated with the test is given by the tail area of the χ^2_ν distribution, with χ^2_ν equal to -2 times the log likelihood ratio, and ν equal to the number of new parameters in the more complicated model. In a Bayesian setting, one would perform such model comparison by calculating a Bayes factor. Although the Bayes factor depends on the sizes of the parameter spaces of the models via the priors for the models' parameters, an approximate and automatic Bayes factor has been found to be useful in practice, at least in the early stages of an analysis. Known both as the Schwarz Criterion and as the Bayesian Information Criterion (BIC), it uses a Gaussian approximation for calculating average likelihoods much as was done above, but additionally uses an "automatic" prior with a width roughly corresponding to the width of the individual data factors in the likelihood. The result is that the log Bayes factor can be approximated as

$$\ln B \approx \ln \left[\mathcal{L}_2(\hat{\theta}, \hat{\phi}) / \mathcal{L}_1(\hat{\theta}) \right] - \frac{1}{2} m_\phi \ln N, \tag{6}$$

where model 2 is the more complicated model, with additional parameters ϕ, and m_ϕ is the dimension of ϕ. The log likelihood ratio is adjusted for the differing number of parameters, but in a way that depends (weakly) on the number of data, a dependence absent from the asymptotic frequentist likelihood ratio test (fixing the inconsistency of that test). Kass and Wasserman (1995) discuss this approximation and the interpretation of the resulting approximate Bayes factors.

4. Methods for Low Dimensional Models

The Laplace approximation is an asymptotic approximation. The remainder of the Bayesian integration methods we will discuss are approximate only in the sense of having accuracies limited by computational resources; they are "exact" methods in the sense of not requiring any approximation of the integrand. We first discuss methods for low dimensional models, where "low" in practice means $m \lesssim 10$ or 20. The successful methods are based on two familiar classes of numerical integration techniques that by themselves are of rather limited utility in Bayesian calculations. It is modification and combination of these basic methods that has led to successful algorithms.

The most familiar approach to numerical integration is the use of quadrature rules—approximating an integral as a weighted sum of values of the integrand. The various rules differ in the choices of weights and abscissas. The impressive thing about these rules in one dimension is that one can take advantage of known "smoothness" properties of the integrand (such as membership in a polynomial family, or approximate similarity to a polynomial times a known nonlinear function such as a Gaussian) in order to construct rules with fast rates of

convergence. Errors falling as $1/n^2$ or $1/n^4$, with n the number of abscissas, are easily achieved. Unfortunately, this desirable behavior is lost when one extends such rules to many dimensions. The simplest approach is to use a product rule—a combination of 1-d rules for each dimension. This suffers from the infamous "curse of dimensionality;" the number of points needed grows exponentially with dimension, so the errors now only fall like $1/n^{2/m}$ or $1/n^{4/m}$. This makes these rules impractical for all but the lowest dimensions ($m \lesssim 4$). Rules exist that extend the quadrature idea to higher dimensions in a more complicated way; they appear under various names in the literature, including cubature, lattice, and monomial rules. In these rules, the abscissas do not lie on a cartesian grid; instead, they are spread over a more complicated multidimensional lattice. But the curse of dimensionality persists for these methods, though it is weakened.

Another familiar approach to numerical integration is Monte Carlo integration—approximating an integral with an average of values of its integrand chosen randomly according to some rule. The great virtue of this approach is that the error falls like $1/\sqrt{n}$, regardless of dimension (or possibly as quickly as $1/n$ if one uses quasirandom sampling rules). This convergence rate is quite poor for one or two dimensions; Monte Carlo integration takes no account of the smoothness information built into quadrature rules and is thus no competition for them in low dimensions. But as soon as $m \gtrsim 3$, the Monte Carlo convergence rate starts to look attractive compared to that of quadrature. Unfortunately, although the *rate* of convergence is attractive, the actual *size* of the errors can be so large as to make implementation impractical. The size of the error depends on the variance of the integrand. This can be reduced in some cases by reweighting the integrand and adjusting the sampling rule to compensate ("importance sampling"). But such reweighting is extremely hard to design in practice unless $m \lesssim 6$.

Although neither of these approaches is useful for Bayesian integration in more than a few dimensions, the successful methods in current use build directly on them. Two families of methods are particularly useful.

First are *randomized quadrature* methods. These methods resemble quadrature rules, but the abscissas are "dithered" randomly. They can combine the virtues of both quadrature and Monte Carlo methods, while diminishing the drawbacks of those methods. The most useful such methods resemble Gaussian quadrature in that they are best used if you can consider the posterior to be reasonably well approximated by, say, a multivariate Gaussian times multinomials (see, e.g., Monahan and Genz 1997).

Second are *subregion-adaptive quadrature* methods. These methods typically use two low-order lattice rules (say, exact for multinomials of order 5 and 7) to estimate an integral *and its error* (via the difference between the estimates from the two rules) in various subregions of the parameter space. The method is applied recursively in regions with the largest error until the entire integral is evaluated to the desired accuracy. These methods automatically put more quadrature points where most of the posterior probability lies, settling for less accuracy in the unimportant tails of the distribution (which often account for a large amount of the parameter space volume).

Randomized and subregion-adaptive methods exist in the general computational literature, but in recent years there has been research effort devoted to tailoring versions of them specifically for Bayesian integration (e.g., Genz

and Kass 1997). The resulting algorithms are not trivial to code from scratch; but fortunately, free and easy-to-use subroutines are readily available. Particularly noteworthy is the work of Alan Genz. The ADAPT subregion-adaptive quadrature algorithm he developed with Malik is available on the World Wide Web; I have used this subroutine in several Bayesian calculations (see, e.g., Loredo, Flanagan, and Wasserman 1997). In addition, his BAYESPACK subroutine package provides a well-documented unified interface to several FORTRAN subroutines for performing Monte Carlo, randomized quadrature, and subregion-adaptive quadrature for Bayesian problems of moderate dimension. Links to this software are available at the BIPS web site previously mentioned.

5. Methods for High Dimensional Models

The majority of straightforward parametric models have a small enough number of parameters that the methods above are likely to prove adequate for implementation of the Bayesian recipe. But sometimes large parametric models must be analyzed (e.g., in analyses of the cosmic background radiation). Also, nonparametric models, once coded, are for all intents and purposes "mega-parametric" models, with each of the discrete sample points or pixels of the estimated function or image playing the role of a parameter. In such problems, the number of parameters can easily reach 10^6 or more. Given the previous discussion, it may seem that a truly Bayesian analysis in such a context, requiring integrals in a highly correlated mega-dimensional space, is simply out of the question. Surprisingly, such analyses have been done routinely for some years now. The basic concept underlying the successful methods is *posterior sampling*, and the most powerful algorithms for implementing them are *Markov Chain Monte Carlo* (MCMC) algorithms.

The notion of posterior sampling should appeal to those used to coding Monte Carlo algorithms in a frequentist context. To calculate a frequentist integral, one can simulate data, as described in § 2, evaluating the integrand for each full data sample. Posterior sampling proceeds similarly, evaluating Bayesian integrals by "simulating hypotheses" rather than data; that is, by drawing samples of θ from the posterior distribution. The great virtue of this approach is that once one has generated a large number of samples θ_i ($i = 1$ to n) from the posterior, that single set of samples can be used to evaluate a whole host of desired integrals for summarizing the implications of the posterior. The marginal distribution for any parameter or subset of parameters can be found simply by ignoring the values of the uninteresting parameters, and plotting or calculating properties of the set of interesting parameter values. Moments of any function of the parameters (including an indicator function whose "moment" gives the probability in a credible region) can be found by averaging the function over the samples. In the case of nonparametric inference, a movie can be made where each frame shows one m-dimensional θ_i (a particular function or image consistent with the data). Viewing this movie will reveal the level of confidence one should have in certain features of the inferred function; if the feature is consistently present in the samples, it has a high probability of being real. Such probabilities can be quantified by calculating appropriate functionals of the samples. This

aspect of the approach is appealing enough that posterior sampling is also worth considering for problems with moderate or even low dimensionality.

The problem is, how can we get the samples in the first place? As already noted, Bayesian integration is complicated precisely because the correlations in parameter space do not make Bayesian integrals amenable to the independent sampling methods underlying frequentist simulation-based calculations.

The simplest method for drawing samples from a complicated probability density function is the well-known *rejection method* (Press et al. 1992, § 7.3). This method depends on the ability to construct a good rejection function that resembles the posterior, but from which we know how to sample efficiently. For problems with dimensions $\lesssim 6$, this is often possible, though it can take a lot of work. But for larger numbers of dimensions, all easily constructed rejection functions waste too much volume, and the overwhelming majority of candidate samples end up being rejected. Thus the rejection method cannot take us into regimes not already accessible by other methods (but it can compete with those methods in the low dimensional regime).

To an audience of physical scientists, a possible solution to the problem appears when it is recast in more suggestive notation. Define a function $\Lambda(\theta)$ according to $\Lambda(\theta) = -\ln[p(\theta|M)\mathcal{L}(\theta)]$. Then the posterior distribution can be written as $p(\theta|D,M) = e^{-\Lambda(\theta)}/Z$, where $Z \equiv \int d\theta\, e^{-\Lambda(\theta)}$. Evaluation of the posterior now resembles two classes of problems familiar to physicists: evaluating Boltzmann factors and partition functions in statistical mechanics, and evaluating Feynman path weights and path integrals in Euclidean quantum field theory. Accordingly, beginning in the mid 1980s, statisticians have mined the computational physics literature for methodology from computational statistical mechanics and quantum field theory on the lattice that could be adapted to evaluate Bayesian integrals. It seems about time this methodology found its way back to the physical sciences.

All of these methods work by constructing a kind of random walk (Markov chain) in the parameter space such that the probability for being in a region of the space is proportional to the posterior density for that region. One starts somewhere (more or less anywhere, though it helps to start at a place of high probability), calculates the posterior density there, takes a small step, and then recalculates the density. The step is accepted or rejected according to some rule, and the process is repeated. The resulting output is a series of points in the sample space, usually discussed and analyzed as if it were a time series. The various methods differ according to the rules used to make the moves in the parameter space, and the rules determining whether or not a step is accepted. It turns out to be extremely simple to invent such rules that are guaranteed to produce accepted steps that correctly sample the posterior. Metropolis et al. (1953) gave the basic algorithm; a discussion of it and some useful modern extensions that is particularly accessible to physical scientists is available in the first section of Toussaint's very readable introduction to methods for lattice QCD calculations (Toussaint 1989), a readable tutorial for statistics students is available in Chib and Greenberg (1995), and a more thorough (but now somewhat dated) review from the point of view of a computer scientist is provided by Neal (1993).

Although the resulting MCMC methods are simple in principle, considerable art is required to implement them in practice. Convergence of MCMC

methods can be guaranteed, but it takes some time for the series of samples to reach equilibrium; and once equilibrated, the samples comprising the output time series are obviously correlated. Part of the art of these MCMC methods is in choosing rules that lead to quick convergence and short correlation lengths. Aspects of the structure of a particular problem, such as availability of conditional distributions for single parameters, or availability of derivatives, can be used to produce effective algorithms for that problem. Once the algorithm is settled upon, one must also take care in determining when the resulting Markov chain has reached equilibrium, and in taking into account the correlations in the resulting samples when calculating inferences. Good output analysis is crucial to successful implementation of MCMC methods and must be done with some sophistication; one has essentially created a new inference problem (albeit a simpler one) in the course of solving the original one.

The development and application of MCMC methods has been the most active area of research in applied Bayesian statistics in recent years. As a result, there is a considerable body of knowledge to draw from when constructing an MCMC algorithm and analyzing its output. A recent volume of review and application papers collects some of this wisdom (Gilks, Richardson, and Spiegelhalter 1996); a roundtable discussion collects further informal advice for novice practitioners (Kass et al. 1998). A monograph discusses engineering applications (Ó Ruanaidh and Fitzgerald 1996), and there is an MCMC web site[3] devoted to publicizing developments in MCMC methods. Although there is not yet a "black box" MCMC algorithm with universal applicability (and there is not likely to be one), good and simple algorithms exist with considerable generality (e.g., for image analysis and density estimation with common priors). In any case, the algorithms themselves are so easy to code that experimentation is relatively straightforward.

Two complications are worth mentioning. First, development of MCMC methods for model comparison lags behind that for parameter estimation. Such methods so far tend to be highly specialized to particular problems (though see Neal 1993 for descriptions of general methods based on algorithms for calculating free energies that have not yet been thoroughly explored in the statistics community). Second, the virtues of posterior sampling may tempt one to apply MCMC methods to parametric models with small numbers of parameters; but some caution is appropriate. MCMC methods so far appear to be most successful when the parameter space is in some sense "homogeneous," with each parameter having a similar role in the likelihood and the same physical dimensions (as with image pixels); otherwise, a "metric" is needed in the parameter space, complicating the analysis.

Asymptotic approximations make the transition to (approximate) Bayesian inference relatively painless for those with existing frequentist codes. Subregion-adaptive quadrature, randomized quadrature, and posterior sampling make more exact Bayesian calculations straightforward for models with significant, realistic complexity. These methods have played a significant role in fueling the explosion of interest in the Bayesian approach in applied statistics. In astrostatistics, the last five years have seen a strong growth of interest in Bayesian inference; these

[3] `http://www.stats.bris.ac.uk/MCMC/`

modern computational tools will certainly play a key role in expanding this interest even further.

References

Bretthorst, G. L. 1988, Bayesian Spectrum Analysis and Parameter Estimation (New York: Springer-Verlag)

Chib, S., & Greenberg, E. 1995, The Amer. Statistician, 49, 327

Evans, M., & Swartz, T. 1995, Statistical Sci., 10, 254

Genz, A., & Kass, R. E. 1997, J. Comp. Graph. Stat., 6, 92

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. 1996, Markov Chain Monte Carlo in Practice (London: Chapman & Hall)

Gregory, P. C., & Loredo, T. J. 1992, ApJ, 398, 146

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. 1998, "Markov chain Monte Carlo in practice: A roundtable discussion," American Statistician, to appear; see MCMC web site

Kass, R. E. and Wasserman, L. 1995, J. Amer. Stat. Assoc., 90, 928

Kass, R. E., Tierney, L., & Kadane, J. B. 1991, in Statistical Multiple Integration, ed. N. Flournoy & R. K. Tsutukawa (Providence, RI: AMS), 89

Kendall, M., & Stuart, A. 1977, The Advanced Theory of Statistics (Londong: Charles Griffin & Company Limited)

Loredo, T. J. 1990, in Maximum Entropy and Bayesian Methods, ed. P. F. Fougere (Dordrecht: Kluwer), 81

Loredo, T. J. 1992, in Statistical Challenges in Modern Astronomy, ed. E. Feigelson & G. Babu (New York: Springer-Verlag), 275

Loredo, T. J., Flanagan, E. E., & Wasserman, I. M. 1997, Phys.Rev.D, 56, 7507

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087

Meyer, S. L. 1975, Data Analysis for Scientists and Engineers (New York: John Wiley & Sons)

Monahan, J., & Genz, A. 1997, J. Amer. Stat. Assoc., 92, 664

Neal, R. 1993, Probabilistic Inference using Markov Chain Monte Carlo Methods, Tech. Report CRG-TR-93-1, Dept. Comp. Sci., U. Toronto; see MCMC web site

O'Hagan, A. 1994, Kendall's Advanced Theory of Statistics, Volume 2B, Bayesian Inference (New York: Halsted Press)

Ó Ruanaidh, J. J. K., & Fitzgerald, W. J. 1996, Numerical Bayesian Methods Applied to Signal Processing (New York: Springer-Verlag)

Press, W. H., Teukolsky, S. A., Flannery, B. P., & Vetterling, W. T. 1992, Numerical Recipes in C: The Art of Scientific Computing (Cambridge University Press)

Sivia, D. S. 1996, Data Analysis: A Bayesian Tutorial (Oxford Univeristy Press)

Tierney, L., & Kadane, J. B. 1986, J. Amer. Stat. Assoc., 81, 82

Toussaint, D. 1989, Comp. Phys. Comm., 56, 69