TI 2002-084/4
Tinbergen Institute Discussion Paper

A Comparison of Marginal Likelihood Computation Methods

Charles S. Bos

Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

A Comparison of Marginal Likelihood Computation Methods

Charles S. Bos *
Faculty of Economics and Operations Research,
Vrije Universiteit *Amsterdam*

August 20th, 2002

Abstract

In a Bayesian analysis, different models can be compared on the basis of the expected or marginal likelihood they attain. Many methods have been devised to compute the marginal likelihood, but simplicity is not the strongest point of most methods. At the same time, the precision of methods is often questionable.

In this paper several methods are presented in a common framework. The explanation of the differences is followed by an application, in which the precision of the methods is tested on a simple regression model where a comparison with analytical results is possible.

JEL classification: C11, C52, C63
Keywords: Marginal likelihood, Bayesian analysis

1 Introduction

In Bayesian inference, there has always been a time gap between the development of new methods and the implementation. The Metropolis-Hastings sampling method dates back to Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953), but only since Chib and Greenberg (1995) is it well understood, and in use, in the econometric society. Likewise when Kloek and Van Dijk (1978) introduced the Importance sampling algorithm, implementation was still rather hard because of the lack of computational power, a situation which has drastically improved.

After estimating the parameters of a model in a Bayesian fashion, it is often of interest to contrast the fit of two competing models. In papers by e.g. Aitkin (1991), Kass and Raftery (1995) and Carlin and Chib (1995), the concepts of marginal likelihood, Bayes factors and posterior odds are explained. Kass and Raftery (1995) even summarizes a range of computational methods for obtaining estimates of the Bayes factor, with an indication of the accuracy that can be expected.

The marginal likelihood promises to provide a convenient method of comparing models by their fit, with less theoretical problems attached to it than encountered when comparing non-nested models in a classical framework. A number of articles (e.g. McCulloch and Rossi 1992, Koop and Potter 1999, Koop and Van Dijk 2000) are appearing where the methods are applied, but no extensive analysis of the practical precision of the methods was performed yet. In this paper, first an overview of the available estimation methods is given in section 2. For each of the methods an indication of the precision is obtained in section 3. The precision is evaluated in a small simulation exercise, on a simple regression model where analytical results are available.

*Correspondence to Charles S. Bos, Faculty of Economics and Operations Research, *Vrije* Universiteit *Amsterdam*, De Boelelaan 1105, NL-1081 HV Amsterdam, The Netherlands. Email: cbos@feweb.vu.nl. Numerous discussions with Herman K. van Dijk, Richard Paap and others are gratefully acknowledged.

2 Marginal likelihood and its computation

2.1 The concept of marginal likelihood and posterior odds

When models are estimated in a classical manner, they can be compared on the basis of the likelihood they attain. The likelihood function is evaluated in the point indicated by the parameter estimates, often at the location of maximum likelihood. In a Bayesian framework, there is not one parameter vector characterizing the fit of the model. Instead, based on the likelihood and the prior, the full posterior distribution of the parameters is derived. Characteristic for the fit of a model M is in this case the expected or marginal likelihood $m(Y \mid M)$, where the expectation is taken over the likelihood $\mathcal{L}(Y;\theta \mid M)$ with respect to the prior distribution $\pi(\theta \mid M)$ of the parameters,

$$m(Y \mid M) = \int_\theta \mathcal{L}(Y;\theta \mid M)\pi(\theta \mid M)\, \partial\theta. \tag{1}$$

The marginal likelihood is the major ingredient for statistics like the Bayes factor BF$= m(Y|M_1)/m(Y|M_2)$, comparing the evidence in favour of two competing models. The posterior odds (PO) ratio is again based on the BF, with PO$=$ BF$\times\pi(M_1)/\pi(M_2)$, and relates the posterior evidence of the models.

Though other ways exist to compute the Bayes factor or the posterior odds (see e.g. Dickey (1971) and Verdinelli and Wasserman (1995) for the (generalized) Savage-Dickey density ratio), the method using the marginal likelihoods is conceptually the simplest. The next section focuses on the computational methods for the marginal likelihoods.

2.2 Computational methods

Only in very special cases, most notably for the exponential likelihood with conjugate priors, the marginal likelihood m can be calculated analytically as the integrating constant of the posterior kernel.[1] In other cases, numerical methods are needed. Table 1 summarizes a range of methods. Details can be found in Kass and Raftery (1995), Chib (1995) and Bos (2001).

The following remarks concerning the methods in table 1 can be made: The brute-force integration method suffers from the curse of dimensionality. When it is not viable anymore, a simulation method may help. The method m_{IS} is not operational without a choice for the importance sampling density $\pi^*(\theta)$, approximating the prior density. Using the prior density as importance function leads to weights $w_i \equiv 1$ as in m_{Prior}, but many drawings will fall in low-likelihood regions. Sampling from the posterior density gives more drawings in the correct region, but leads to an estimate of m which may not have a finite variance (Newton and Raftery 1994). Intermediate positions, like $\pi^*(\theta) = \delta\pi(\theta) + (1-\delta)p(\theta|Y)$ can be chosen: This gives a consistent estimate with better convergence behaviour (see also Newton and Raftery 1994). A more recent solution for stabilizing the harmonic mean estimate, utilizing a technique of lowering the dimension of the problem, is given in Satagopan, Newton and Raftery (2000), and is not discussed here.

m_{LP} and m_{Kern} are special versions of m_{App}. The first essentially fits a normal density to the mode of the posterior. The method can be expected to work well in cases where the posterior is highly peaked, less well in multimodal cases. m_{Kern} applies a kernel smoother, and therefore also suffers to some extent from the curse of dimensionality, as more drawings $\theta^{(i)}$ are needed when the dimension increases, in order to get a good approximation to the posterior density.

Finally, the Gibbs sampling method can be applied in cases where the conditional densities are available, and may be the only method when data augmentation (Gelfand and Smith 1990) is used. Implementing may prove to be hard, and for each element of the vector of parameters a separate Gibbs chain has to be run, leading to a rather hefty computational burden, as compared to the other methods.

[1] As we are calculating the marginal likelihood of a specific model M on a fixed data set Y in this section, dependence of m on these quantities is suppressed in the notation.

Table 1: Computational methods for marginal likelihoods

$m_{\text{Anal}} = \int \mathcal{L}(Y,\theta)\pi(\theta)\partial\theta$
The exact solution, available only in special cases.

m_{Num} Using brute-force numerical integration, m can be computed for low-dimensional problems.

$m_{\text{Prior}} = \mathsf{E}_{\pi(\theta)}\mathcal{L}(Y,\theta) \approx \frac{1}{n}\sum \mathcal{L}(Y,\theta^{(i)}); \theta^{(i)} \sim \pi(\theta)$
Simulating from the prior, m is the average likelihood. A very unstable and inefficient estimate.

$m_{\text{IS}} = \frac{1}{\sum w_i}\sum w_i \mathcal{L}(Y,\theta^{(i)}); \theta^{(i)} \sim \pi^*(\theta), w_i = \pi(\theta^{(i)})/\pi^*(\theta^{(i)})$
Sampling from an importance density with higher probability mass around the posterior mode improves on m_{Prior}.

$m_{\text{HM}} = \left(\frac{1}{N}\sum \frac{1}{\mathcal{L}(Y,\theta^{(i)})}\right)^{-1}; \theta^{(i)} \sim p(\theta|Y)$
Using the posterior density as importance sampling density in m_{IS} leads to using the harmonic mean of the likelihood values as estimator.

$m_{\text{App}} = \mathcal{L}(Y,\theta)\pi(\theta)/p(\theta|Y)$
In general, relating the height of the posterior kernel to the height of an approximating posterior density in a high density point θ can give an estimate of m.

$m_{\text{LP}} = (2\pi)^{k/2}|\tilde{\Sigma}|^{\frac{1}{2}}\mathcal{L}(Y,\tilde{\theta})\pi(\tilde{\theta})$
This is m_{App} using a quadratic expansion of the posterior kernel around the posterior mode $\tilde{\theta}$.

$m_{\text{Kern}} = \mathcal{L}(Y,\theta)\pi(\theta)/p_{\text{Kern}}(\theta|Y)$
This is m_{App} using a kernel smoother to approximate the posterior density at a high density point θ.

$m_{\text{Gibbs}} = \mathcal{L}(Y,\theta)\pi(\theta)/\left[p(\theta_1|Y)\prod p(\theta_i|\theta_1,..,\theta_{i-1},Y)\right]$
Chib (1995) proposed to use $k-1$ samples from separate Gibbs chains to approximate $p(\theta_i|\theta_1,..,\theta_{i-1},Y) \approx \sum p(\theta_i|\theta_1,..,\theta_{i-1},\theta_{i+1}^{(j)},..,\theta_k^{(j)},Y)/N$, where N samples are drawn for the missing elements $\theta_{i+1},..,\theta_k$ from the conditional densities.

3 Marginal likelihood in practice: A comparison

3.1 Data, model and analytical marginal likelihood

In Brownlee (1965, page 454) a data-set concerning the oxidation of ammonia for producing nitric acid is examined. Explanatory variables are the air flow in the tower, the temperature of cooling water and the acid concentration which is produced. The dependent variable is the so-called stack-loss, which is 10 times the percentage of ammonia which escapes from the process without having been converted into nitric acid.

This data-set was analysed often (Atkinson 1985, Hoeting, Raftery and Madigan 1996, Justel and Peña 1996) using regression models, with the focus on recognizing outliers among the 21 observations on the 4 variables. For the purpose of displaying the differences of the computational methods for marginal likelihoods, it is sufficient to limit ourselves to the pure regression model

$$y = X\beta + \epsilon, \qquad \epsilon = \mathcal{N}(0,\sigma^2\mathcal{I}_n). \tag{2}$$

For simplicity, we assume that the variance σ^2 is known, and fix it at its least squares estimate.

With a normal prior $\pi(\beta)$, with expectation β_0 and covariance matrix $\Sigma_0 = \sigma_0^2\mathcal{I}_k$, the posterior $p(\beta|Y)$ is normal as well, with covariance $\tilde{\Sigma} = (\hat{\Sigma}^{-1} + \Sigma_0^{-1})^{-1}$ and expectation $\tilde{\beta} = \tilde{\Sigma}^{-1}(\hat{\Sigma}^{-1}\hat{\beta} + \Sigma_0^{-1}\beta_0)$, and $\hat{\beta}$ and $\hat{\Sigma}$ the least squares estimates. Computing the marginal likelihood from the quotient of the posterior kernel $\mathcal{L}(Y;\theta)\pi(\theta)$ and the posterior density $p(\theta|Y)$ gives the expression

$$m_{\text{Anal}} = (2\pi)^{-\frac{n}{2}}\sigma^{-n}|\Sigma_0|^{-\frac{1}{2}}|\tilde{\Sigma}|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left[\frac{y'y}{\sigma^2} + \beta_0\Sigma_0^{-1}\beta_0 - \tilde{\beta}\tilde{\Sigma}^{-1}\tilde{\beta}\right]\right) \tag{3}$$

Note how all terms containing β drop out of this equation, as should be the case for the integrand of the posterior kernel over β.

3.2 Estimation results

Table 2 displays the main estimation results for the regression model on the stack-loss data. For the prior, an expectation of zero and a large standard deviation of the elements was chosen, such that the prior hardly influences the results. Indeed, the difference between the least squares and posterior estimates is minimal.

Table 2: Prior, least squares and posterior moments of the parameters

	Prior		Least squares		Posterior	
Variable	β_0	σ_0	$\hat{\beta}$	$\hat{\sigma}$	$\tilde{\beta}$	$\tilde{\sigma}$
Air flow	0	20	0.7968	0.166	0.7969	0.166
Water temperature	0	20	1.1114	0.456	1.1109	0.456
Acid concentration	0	20	−0.6250	0.085	−0.6249	0.085

Note: Non-zero correlations between the LS and posterior parameters are not reported.

Table 3: Logarithm of the marginal likelihood of the regression model for Stack-loss data

		Repeated simulation				
Method	Log m	Mean	s.d.	Min	Max	%Time
m_{Anal}	−74.271					
m_{Num}	−74.271					
m_{Prior}	−110.710	−759.827	558.03	−3141.87	−81.90	11
m_{HM}	−61.281	−61.111	0.57	−64.18	−60.16	12
m_{LP}	−74.271	−74.280	0.04	−74.38	−74.18	0
m_{Kern}	−74.166	−74.109	0.08	−74.30	−73.80	1
m_{Gibbs}	−74.354	−74.304	0.49	−75.30	−72.92	76

Note: Results are reported for one large simulation of 50.000 drawings and for 100 repeated simulations of sample size 1.000. Timings indicate the percentage of time spent computing the measures in the repeated simulation.

The methods explained in section 2 were used for constructing the figures in table 3. For each of the methods indicated in the first column, the table reports an estimate of the logarithm of marginal likelihood in the second column.

With the numerical integration, the result of the analytical computation can be found if enough time/effort is put into the integration routine. For this model, no difference is found indeed. However in practice, for more elaborate models, it is often difficult to find sensible bounds for the integration region, and either the tails of the posterior density might be missed or too much effort is put into regions with little mass, leading to large computational efforts being wasted.

Both the m_{Prior} and m_{HM} methods are special versions of m_{IS}. They both suffer from instability: Even after more than 45.000 drawings,[2] the estimates of $\log m_{\text{Prior}}$ and $\log m_{\text{HM}}$ can be seen to change.

The LaPlace method uses a normal density to approximate the normal posterior density: As the approximation is perfect, the correct solution is found. In general cases the estimate will not be perfect, but practice indicates that this method works rather well, especially for highly peaked, unimodal posterior densities.

[2]For m_{Prior}, 50.000 values are sampled from the prior density. For m_{HM}, and likewise for m_{Kern}, a sample of 50.000 *accepted* drawings from a Metropolis-Hastings chain with a multivariate Student-t candidate with 4 degrees-of-freedom was collected. At an acceptance rate of 73%, this gave a total sample of size 68.071 from the posterior density.

Approximating the posterior with a kernel smoothing density (a multivariate Gaussian kernel smoother with automatic bandwidth selection was used) is simple, but gives good results as well. The Gibbs method gives an answer which is slightly more precise, but is more costly in terms of computing time and in complications in the estimation procedure.

Columns 3–7 of table 3 report results for replicating the estimation 100 times, using sample sizes of 1.000. The prior method proves to be totally unreliable, with an enormous standard deviation in column 4. The harmonic mean method does not converge to the correct estimate either.[3] For the LaPlace method, the estimate was now computed at the mean of the parameters sampled from the posterior density, with the covariance of the sample as covariance estimate. It is seen that this very simple method is quite good (at least for this model). Also the kernel and Gibbs methods perform well, with more variation in results for the latter.

The last column gives an indication of the respective durations of the calculations. The m_{Prior} and m_{HM} methods use rather costly function evaluations, leading to computation times increasing with the size of the sample. The method applying Gibbs sampling needs to collect new samples from the conditional densities, through several runs of the Gibbs chain. In the present implementation it is taking about 7 times as long as either m_{Prior} or m_{HM}. Depending on the sample size, the dimension of the problem and the integration method used, it can be quicker than the brute-force integration method. For the results in column 2, on the larger sample, the Gibbs computation finished in roughly one third of the time needed for computing the integral using Gaussian quadrature.

4 Concluding remarks

In this paper, a range of computational methods for obtaining an estimate of the marginal likelihood of a model are reviewed shortly, and applied in a small application. This way, the qualitative impression of the precision of the methods available until now was checked in a quantitative manner.

The main finding is that both the method of computing the marginal likelihood as the expectation of the likelihood under the prior, as the method of computing the harmonic mean of likelihood values when sampling from the posterior, are not trustworthy. Though work is under way to stabilize m_{HM} (Satagopan et al. 2000), the methods applying either a LaPlace or kernel approximation to the posterior density in a high-density location prove to be more useful. The Gibbs sampling method of Chib (1995) can give good results as well, though the computational costs are high especially with increasing dimension of the vector of parameters.

References

Aitkin, M. (1991), 'Posterior Bayes factors', *Journal of the Royal Statistical Society, Series B* **53**(1), 111–142.

Atkinson, A. C. (1985), *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford statistical science series, Clarendon, Oxford.

Bos, C. S. (2001), Time Varying Parameter Models for Inflation and Exchange Rates, PhD thesis, Tinbergen Institute, Erasmus University Rotterdam. TI 256.

[3] A separate computation, continuing for 10^9 drawings from the posterior, continued to display jumps in the cumulative estimate of m_{HM} even at the end of the run.

Brownlee, K. A. (1965), *Statistical Theory and Methodology in Science and Engineering*, 2 edn, Wiley, New York.

Carlin, B. P. and Chib, S. (1995), 'Bayesian model choice via Markov chain Monte Carlo methods', *Journal of the Royal Statistical Society, Series B* **57**(3), 473–484.

Chib, S. (1995), 'Marginal likelihood from the Gibbs output', *Journal of the American Statistical Association* **90**(432), 1313–1321.

Chib, S. and Greenberg, E. (1995), 'Understanding the Metropolis-Hastings algorithm', *The American Statistician* **49**(4), 327–335.

Dickey, J. (1971), 'The weighted likelihood ratio, linear hypotheses on normal location parameters', *Annals of Statistics* **42**, 204–224.

Gelfand, A. E. and Smith, A. F. M. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**(410), 398–409.

Hoeting, J. A., Raftery, A. E. and Madigan, D. (1996), 'A method for simultaneous variable selection and outlier identification in linear regression', *Computational Statistics and Data Analysis* **22**, 251–270.

Justel, A. and Peña, D. (1996), 'Gibbs sampling will fail in outlier problems with strong masking', *Journal of Computational & Graphical Statistics* **5**(2), 176–189.

Kass, R. E. and Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Kloek, T. and Van Dijk, H. K. (1978), 'Bayesian estimates of equation system parameters: An application of integration by Monte Carlo', *Econometrica* **46**, 1–20.

Koop, G. and Potter, S. M. (1999), 'Bayes factors and nonlinearity: Evidence from economic time series', *Journal of Econometrics* **88**, 251–281.

Koop, G. and Van Dijk, H. K. (2000), 'Testing for integration using evolving trend and seasonal models: A Bayesian approach', *Journal of Econometrics* **97**(2), 261–291.

McCulloch, R. E. and Rossi, P. (1992), 'Bayes factors for nonlinear hypotheses and likelihood distributions', *Biometrika* **79**, 663–676.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equations of state calculations by fast computing machines', *Journal of Chemical Physics* **21**, 1087–1091.

Newton, M. A. and Raftery, A. E. (1994), 'Approximate Bayesian inference by the weighted likelihood bootstrap', *Journal of the Royal Statistical Society, Series B* **3**, 3–48.

Satagopan, J. M., Newton, M. A. and Raftery, A. E. (2000), Easy estimation of normalizing constants and Bayes factors from posterior simulation: Stabilizing the harmonic mean estimator. Unpublished manuscript.

Verdinelli, I. and Wasserman, L. (1995), 'Computing Bayes factors using a generalization of the Savage-Dickey density ratio', *Journal of the American Statistical Association* **90**(430), 614–618.