EC 607: Bayesian Econometrics

Bayesian Computation III

Prof. Jeremy M. Piger

Department of Economics

University of Oregon

Last Revised: March 22, 2019

1 Methods to Estimate the Marginal Likelihood via Simulation

- Recall, the marginal likelihood is a crucial element of the Bayesian approach to comparing alternative models.
- The marginal likelihood is given by:

$$\begin{aligned} p(Y) &= \int_{\theta} p(Y,\theta)\, d\theta \\ &= \int_{\theta} p(Y|\theta)\, p(\theta)\, d\theta \\ &= E_{p(\theta)}\left(p(Y|\theta)\right) \end{aligned}$$

- We normally can't evaluate this integral/expectation analytically. Here we will discuss approaches to estimate this integral via simulation.

1.1 What Not to Do: Direct Monte Carlo Integration Using Simulation from the Prior

- The marginal likelihood is just an expected value of a function taken with respect to the prior density function. Thus, simulation via Monte Carlo Integration provides a (seemingly) obvious approach to estimate the marginal likelihood. In particular, suppose we have G draws of θ from the **prior** distribution for θ, $p(\theta)$. Such draws are generally easy to construct as the prior distribution for θ is usually a known distribution. We could then approximate $p(Y) = E_{p(\theta)}\left(p(Y|\theta)\right)$ as:

$$\frac{1}{G}\sum_{g=1}^{G} p\left(Y|\theta^{[g]}\right) \tag{1}$$

However, while obvious, when the prior is even moderately diffuse, this will be a very

inefficient approach to obtaining the marginal likelihood. Specifically, G will need to be very (impossibly) large before a reasonable level of convergence is reached.

1.2 Monte Carlo Integration using Simulation from the Posterior: The Chib (1995) and Chib and Jeliaskov (2001) Methods

- Using Bayes Rule, the marginal likelihood can be written as:

$$p(Y) = \frac{p(Y|\theta)\, p(\theta)}{p(\theta|Y)}$$

This is known as the **Basic Marginal Likelihood Identity**. It is an identity because this equation holds for any value of θ. Define some particular value of θ as $\widetilde{\theta}$. Then:

$$p(Y) = \frac{p\left(Y|\widetilde{\theta}\right) p\left(\widetilde{\theta}\right)}{p\left(\widetilde{\theta}|Y\right)}$$

- Note that we can usually evaluate the likelihood function $p\left(Y|\widetilde{\theta}\right)$ as well as the prior density $p\left(\widetilde{\theta}\right)$ analytically. So, all we need to do calculate the marginal likelihood is to calculate the posterior density at the point $\widetilde{\theta}$.

- This turns the problem of estimating the marginal likelihood into one of estimating the **posterior ordinate** $p\left(\widetilde{\theta}|Y\right)$. We will discuss how this can be done, using only samples from the posterior distribution. We will consider two cases: one where the samples are obtained from the Gibbs Sampler, and the other where the samples are obtained from the Metropolis-Hastings sampler.

Estimating $p\left(\widetilde{\theta}|Y\right)$ using samples from the Gibbs Sampler

The following ideas were presented in Chib(1995, *JASA*). Suppose we have a Gibbs Sampler with three blocks: $\theta = (\theta_1', \theta_2', \theta_3)'$, and have used this Gibbs sampler to

produce G draws of θ from $p\left(\theta|Y\right)$.

The posterior ordinate can be factored as:

$$p\left(\widetilde{\theta}|Y\right) = p\left(\widetilde{\theta}_1|\widetilde{\theta}_2, \widetilde{\theta}_3, Y\right) p\left(\widetilde{\theta}_2|\widetilde{\theta}_3, Y\right) p\left(\widetilde{\theta}_3|Y\right)$$

In a Gibbs Sampler, the conditional posterior distributions are known. That is, we can evaluate:

1. $p\left(\theta_1|\theta_2, \theta_3, Y\right)$
2. $p\left(\theta_2|\theta_1, \theta_3, Y\right)$
3. $p\left(\theta_3|\theta_1, \theta_2, Y\right)$

Thus, the term $p\left(\widetilde{\theta}_1|\widetilde{\theta}_2, \widetilde{\theta}_3, Y\right)$ can be calculated analytically. Next, consider the third term:

$$p\left(\widetilde{\theta}_3|Y\right) = \int_{\theta_2}\int_{\theta_1} p\left(\widetilde{\theta}_3|\theta_1, \theta_2, Y\right) p\left(\theta_1, \theta_2|Y\right) d\theta_1 d\theta_2$$

$$p\left(\widetilde{\theta}_3|Y\right) = E_{p(\theta_1,\theta_2|Y)}\left(p\left(\widetilde{\theta}_3|\theta_1, \theta_2, Y\right)\right)$$

This can be estimated via Monte Carlo integration as:

$$\frac{1}{G}\sum_{g=1}^{G} p\left(\widetilde{\theta}_3|\theta_1^{[g]}, \theta_2^{[g]}, Y\right) \tag{2}$$

where $\theta_1^{[g]}$ and $\theta_2^{[g]}$ are random draws from $p\left(\theta_1, \theta_2|Y\right)$. Such draws are given to us by the draws of θ_1 and θ_2 produced by the Gibbs Sampler.

This leaves us needing to estimate the second term:

$$p\left(\widetilde{\theta}_2|\widetilde{\theta}_3, Y\right) = \int_{\theta_1} p\left(\widetilde{\theta}_2, \theta_1|\widetilde{\theta}_3, Y\right) d\theta_1$$

$$p\left(\widetilde{\theta}_2|\widetilde{\theta}_3, Y\right) = \int_{\theta_1} p\left(\widetilde{\theta}_2|\theta_1, \widetilde{\theta}_3, Y\right) p\left(\theta_1|\widetilde{\theta}_3, Y\right) d\theta_1$$

$$p\left(\widetilde{\theta}_2|\widetilde{\theta}_3, Y\right) = E_{p\left(\theta_1|\widetilde{\theta}_3, Y\right)}\left(p\left(\widetilde{\theta}_2|\theta_1, \widetilde{\theta}_3, Y\right)\right)$$

This can be estimated via Monte Carlo Integration as:

$$\frac{1}{G}\sum_{g=1}^{G} p\left(\widetilde{\theta}_2|\theta_1^{[g]}, \widetilde{\theta}_3, Y\right) \tag{3}$$

where $\theta_1^{[g]}$ are random draws from $p\left(\theta_1|\widetilde{\theta}_3, Y\right)$. Such draws can be produced by running the following **reduced** Gibbs Sampler:

1. Draw $\theta_1^{[g+1]}$ from $p\left(\theta_1|\theta_2^{[g]}, \widetilde{\theta}_3, Y\right)$
2. Draw $\theta_2^{[g+1]}$ from $p\left(\theta_2|\theta_1^{[g+1]}, \widetilde{\theta}_3, Y\right)$

This reduced Gibbs Sampler will produce draws of θ_1 and θ_2 from $p\left(\theta_1, \theta_2|\widetilde{\theta}_3, Y\right)$. The draws of θ_1 will also represent draws from $p\left(\theta_1|\widetilde{\theta}_3, Y\right)$.

A full Gibbs sampler is easy to modify to produce a reduced Gibbs Sampler. We would just remove the step where θ_3 is drawn from the full Gibbs sampler, and replace it with $\theta_3 = \widetilde{\theta}_3$.

To summarize, to compute the marginal likelihood of a model with $\theta = (\theta_1', \theta_2', \theta_3')'$, the posterior of which has been sampled G times by the Gibbs Sampler, we would take the following steps:

1. Pick a value for $\widetilde{\theta} = \left(\widetilde{\theta}_1, \widetilde{\theta}_2', \widetilde{\theta}_3'\right)'$. Although technically any value will do, Chib (1995) suggests using a high density point in the posterior.
2. Evaluate $p\left(Y|\widetilde{\theta}\right)$, $p\left(\widetilde{\theta}\right)$, and $p\left(\widetilde{\theta}_3|\widetilde{\theta}_2, \widetilde{\theta}_1\right)$ analytically.

3. Draw $\theta_1^{[g],\text{reduced}}$, $g = 1, 2, \cdots, G$ from $p\left(\theta_1|\widetilde{\theta}_3, Y\right)$ using a reduced Gibbs Sampler. Then estimate $p\left(\widetilde{\theta}_2|\widetilde{\theta}_3, Y\right)$ as:

$$\widehat{p\left(\widetilde{\theta}_2|\widetilde{\theta}_3, Y\right)} = \frac{1}{G}\sum_{g=1}^{G} p\left(\widetilde{\theta}_2|\theta_1^{[g],\text{reduced}}, \widetilde{\theta}_3, Y\right) \tag{4}$$

4. Estimate $p\left(\widetilde{\theta}_3|Y\right)$ as:

$$\widehat{p\left(\widetilde{\theta}_3|Y\right)} = \frac{1}{G}\sum_{g=1}^{G} p\left(\widetilde{\theta}_3|\theta_1^{[g]}, \theta_2^{[g]}, Y\right) \tag{5}$$

   where $\theta_1^{[g]}$ and $\theta_2^{[g]}$ are draws from the full Gibbs sampler.

5. Form the estimated marginal likelihood:

$$\widehat{p\left(Y\right)} = \frac{p\left(Y|\widetilde{\theta}\right)p\left(\widetilde{\theta}\right)}{p\left(\widetilde{\theta}_1|\widetilde{\theta}_2, \widetilde{\theta}_3, Y\right)\widehat{p\left(\widetilde{\theta}_2|\widetilde{\theta}_3, Y\right)}\widehat{p\left(\widetilde{\theta}_3|Y\right)}}$$

- This procedure can be easily modified to handle a different number of blocks than 3. Additional blocks will mean additional reduced Gibbs runs. Two blocks will not need any reduced Gibbs runs.

- It can also be modified to allow for a Gibbs Sampler that includes data augmentation. In this case, the data augmentation would just be thrown in as an extra step in all the reduced and full Gibbs samplers.

- Note that in this procedure, the Monte Carlo integration is with respect to the posterior distribution, not the prior distribution as in the direct Monte Carlo integration approach. This is why this approach works better than Monte Carlo integration, as the posterior is usually a much tighter density than the prior, and thus can be numerically integrated much more efficiently.

Estimating $p\left(\widetilde{\theta}|Y\right)$ using samples from the Metropolis-Hastings Sampler

The following ideas were presented in Chib and Jeliaskov(2001, *JASA*). Suppose we have sampled θ G times from the posterior density $p\left(\widetilde{\theta}|Y\right)$ using the Metropolis-Hastings algorithm. Here we will focus on the case where the Metropolis-Hastings algorithm has been implemented in one block, but the technique we will discuss can be extended to multiple blocks.

Recall, the MH acceptance probability is given by $\alpha\left(\theta^{[g]}, \theta^{*}\right)$, while the proposal is given by $q\left(\theta^{*}|\theta^{[g]}\right)$. Consider a proposed value, $\theta^{*} = \widetilde{\theta}$. Then, since the MH transition kernel is reversible with stationary distribution $p\left(\theta|Y\right)$, we have:

$$\alpha\left(\theta, \widetilde{\theta}\right) q\left(\widetilde{\theta}|\theta\right) p\left(\theta|Y\right) = \alpha\left(\widetilde{\theta}, \theta\right) q\left(\theta|\widetilde{\theta}\right) p\left(\widetilde{\theta}|Y\right)$$

Integrating both sides of this expression with regards to θ yields

$$\int_{\theta} \alpha\left(\theta, \widetilde{\theta}\right) q\left(\widetilde{\theta}|\theta\right) p\left(\theta|Y\right) d\theta = p\left(\widetilde{\theta}|Y\right) \int_{\theta} \alpha\left(\widetilde{\theta}, \theta\right) q\left(\theta|\widetilde{\theta}\right) d\theta$$

So:

$$p\left(\widetilde{\theta}|Y\right) = \frac{\int_{\theta} \alpha\left(\theta, \widetilde{\theta}\right) q\left(\widetilde{\theta}|\theta\right) p\left(\theta|Y\right) d\theta}{\int_{\theta} \alpha\left(\widetilde{\theta}, \theta\right) q\left(\theta|\widetilde{\theta}\right) d\theta}$$

The numerator of this equation is:

$$E_{p(\theta|Y)}\left[\alpha\left(\theta, \widetilde{\theta}\right) q\left(\widetilde{\theta}|\theta\right)\right]$$

and can be estimated via Monte Carlo integration as:

$$\frac{1}{G} \sum_{g=1}^{G} \alpha\left(\theta^{[g]}, \widetilde{\theta}\right) q\left(\widetilde{\theta}|\theta^{[g]}\right)$$

where $\theta^{[g]}$ are the draws from $p\left(\theta|Y\right)$ produced by the MH algorithm.

The denominator of this equation is:

$$E_{q\left(\theta|\widetilde{\theta}\right)}\left[\alpha\left(\widetilde{\theta},\theta\right)\right]$$

and can be estimated via Monte Carlo integration as:

$$\frac{1}{J}\sum_{j=1}^{J}\alpha\left(\widetilde{\theta},\theta^{[j]}\right)$$

where $\theta^{[j]}$ are draws from $q\left(\theta|\widetilde{\theta}\right)$. These can usually be taken very quickly, as the proposal is easy to sample from.

We then have an estimate of $p\left(\widetilde{\theta}|Y\right)$ as:

$$\widehat{p\left(\widetilde{\theta}|Y\right)}=\frac{\frac{1}{G}\sum_{g=1}^{G}\alpha\left(\theta^{[g]},\widetilde{\theta}\right)q\left(\widetilde{\theta}|\theta^{[g]}\right)}{\frac{1}{J}\sum_{j=1}^{J}\alpha\left(\widetilde{\theta},\theta^{[j]}\right)}$$

- The estimated marginal likelihood is then:

$$\widehat{p\left(Y\right)}=\frac{p\left(Y|\widetilde{\theta}\right)p\left(\widetilde{\theta}\right)}{\widehat{p\left(\widetilde{\theta}|Y\right)}}$$

1.3 Monte Carlo Integration using Simulation from the Posterior: The Modified Harmonic Mean Approach

- Manipulation of Bayes Rule reveals:

$$\frac{1}{p\left(Y\right)}=\frac{1}{p\left(Y|\theta\right)p\left(\theta\right)}p\left(\theta|Y\right)$$

- Consider a pdf, $f\left(\theta\right)$, where $f\left(\theta\right)$ has support contained in the region where $p\left(\theta\right)$,

$p(\theta|Y)$ and $p(Y|\theta)$ are defined. Multiplying both sides of the equation by $f(\theta)$ and integrating, we then have:

$$\frac{1}{p(Y)} = \int_{\theta} \frac{f(\theta)}{p(Y|\theta)\, p(\theta)} p(\theta|Y)\, d\theta$$

This follows because $\int_{\theta} f(\theta)\, d\theta = 1$.

- This suggests a monte carlo integration approach to estimate the marginal likelihood, known as the Modified Harmonic Mean Estimator (MHME):

$$\widehat{p(Y)} = \left[\frac{1}{G} \sum_{g=1}^{G} \frac{f\left(\theta^{[g]}\right)}{p(Y|\theta^{[g]})\, p(\theta^{[g]})} \right]^{-1}$$

where $\theta^{[g]}$ is a draw from $p(\theta|Y)$.

- MHME This provides a very general approach to estimate the marginal likelihood. Provided we have a posterior simulator (such as the Gibbs Sampler or the Metropolis-Hastings algorithm), and we can compute the likelihood function and the prior, then we are able to use this approach to compute the marginal likelihood. Note that this equation requires that we use the full equations for the prior density, the likelihood function, and $f(\theta)$, not just the kernels. This is because constant terms do not cancel out in the numerator and denominator of the ratio above.

- Unfortunately, convergence of the MHME can be problematic. For a poorly chosen value of $f(\theta)$, the Modified Harmonic Mean estimator can have infinite variance. The intuition is that some draws of θ can yield very small likelihood values, which will cause the term in brackets above to explode. The practical upshot is that different runs of the MHME may give you very different answers, even if the number of draws is very large.

- Getting the MHME to work well comes down to carefully selecting $f(\theta)$. Formally, we

need the ratio in brackets above to be bounded from above. Geweke (1999, *Econometric Reviews*) provides one approach that seems to work well in practice. Geweke's idea is to set $f(\theta)$ equal to a large sample approximation to $p(\theta|Y)$ with its tails chopped off. Specifically Geweke proposes:

$$\begin{aligned} f(\theta) &= (2\pi)^{-\frac{k}{2}} |\widehat{V}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}\left(\theta-\widehat{\theta}\right)'\left(\widehat{V}\right)^{-1}\left(\theta-\widehat{\theta}\right)\right] \\ &\quad \times I\left\{\left(\theta-\widehat{\theta}\right)'\left(\widehat{V}\right)^{-1}\left(\theta-\widehat{\theta}\right) \leq F^{-1}_{\chi^2_k}(\tau)\right\}\tau^{-1} \end{aligned}$$

Here, $\widehat{\theta}$ is the posterior mean computed from draws from the posterior. $F^{-1}_{\chi^2_k}$ is the inverse CDF of a chi-squared distribution with k degrees of freedom (k being the number of elements in θ), thus $F^{-1}_{\chi^2_k}(\tau)$ is the τ percentile of the Chi-squared distribution with k degrees of freedom. Finally, $I(\cdot)$ is an indicator function that is 1 if the argument is true and zero otherwise.

- Geweke's propsed $f(\theta)$ is interpreted as follows: The first line is an equation for a multivariate normal density with mean equal to the posterior mean and variance covariance matrix equal to the variance covariance matrix of the posterior. This can be interpreted as a large sample approximation to the posterior density. The second term puts zero weight on extreme observations, by chopping off the tails of this density, where extreme is defined by the percentile τ, where $0 < \tau < 1$. Lower values of τ chop off more of the tails of the normal density. The third term (τ^{-1}) rescales the density such that it integrates to one.

- The intuition for this choice of $f(\theta)$ is that it zeros out low posterior density points, which would have low likelihood values, thus preventing these samples from exploding the ratio in brackets in the MHME equation. Here, τ serves as a tuning parameter, which can be adjusted to assess the stability of the MHME estimator, and thus help us feel more comfortable about convergence. In practice, when the posterior sampler

is run for estimation purposes, the values of $p\left(Y|\theta^{[g]}\right)p\left(\theta^{[g]}\right)$ can be saved, making evaluation of the MHME for alternative values of τ very quick.

2 Exercise: Marginal Likelihood for Linear Regression Model

- Consider the unrestricted linear regression model for course evaluations that you worked with in Week 4. In the remainder of this exercise, I will refer to this model as "Model 1." Recall, you used the Gibbs Sampler to estimate Model 1. In this question you will use the Chib (1995, *JASA*) approach to estimate the marginal likelihood for Model 1 (and related models) using output from the Gibbs Sampler.

- For the linear regression model, $\theta = (\beta, h)$ and the basic marginal likelihood identity becomes:

$$p(Y) = \frac{p\left(Y|\widetilde{\beta},\widetilde{h}\right)p\left(\widetilde{\beta},\widetilde{h}\right)}{p\left(\widetilde{\beta},\widetilde{h}|Y\right)}$$

- The numerator of this equation can be evaluated analytically, since $p\left(Y|\widetilde{\beta},\widetilde{h}\right)$ and $p\left(\widetilde{\beta},\widetilde{h}\right)$ are the likelihood function and prior distribution respectively, evaluated at $\widetilde{\beta}$ and $\widetilde{h}$. Both of these are known in this problem.

- The denominator can be factored using the law of total probability as:

$$p\left(\widetilde{\beta},\widetilde{h}|Y\right) = p\left(\widetilde{\beta}|\widetilde{h},Y\right)p\left(\widetilde{h}|Y\right)$$

- The term $p\left(\widetilde{\beta}|\widetilde{h},Y\right)$ can be evaluated analytically, since this conditional posterior distribution is known when we apply the Gibbs Sampler to the linear regression model. The final term, $p\left(\widetilde{h}|Y\right)$, can be evaluated using Monte Carlo Integration:

$$\frac{1}{G}\sum_{g=1}^{G} p\left(\widetilde{h}|\beta^{[g]}, Y\right) \tag{6}$$

- where $\beta^{[g]}$ is the g^{th} draw from $p(\beta|Y)$ obtained using the Gibbs Sampler. Note in doing this that you should only use post-convergence draws of β.

- In the dropbox you will find a zipped collection of files called "Linear Regression GS with Marginal Likelihood.zip." Work with these files to make sure you understand what they do and how they work. Note that these programs estimate the marginal likelihood using both the method of Chib (1995) and direct monte carlo integration from the prior density. The direct monte carlo estimate is done just to show you how bad it is - use the marginal likelihood estimate from the Chib (1995) approach in any analysis you do.

- Suggested exercises include:

    1. Consider two models, Model 1 (which includes all variables), and Model 2, which excludes the "beauty" variable. Report posterior model probabilities for these two models. How did you set the prior model probabilities?

    2. Consider two models, Model 1 (which includes all variables), and Model 3, which excludes the "Female" variable. Report posterior model probabilities for these two models.

    3. The program sets $\widetilde{\beta}$ and $\widetilde{h}$ equal to the posterior median of these variables. Changes these to be something else. Is the estimated marginal likelihood affected?

    4. How sensitive is the estimate of the marginal likelihood to changing the number of simulations?

    5. Run the program a few times and notice that the direct monte carlo estimate is

very unstable. Can you make it better by doing more simulations? You can play with this by changing the "num_direct_sim" variable.

3 Exercise: Marginal Likelihood for Nonlinear Regression Model

- Consider the nonlinear regression model that you worked with in Week 3. In this question you will use both the Chib and Jeliaskov (2001, *JASA*) and Modified Harmonic Mean Estimator of Geweke(1999, *Econometric Reviews*) to estimate the marginal likelihood for this model.

- The zipped collection of files called "Nonlinear_Regression_MH_with_Marginal_Likelihood.zip" contains Matlab code to estimate the marginal likelihood for a nonlinear regression model. Work with these files to make sure you understand what they do and how they work. Suggested exercises include:

    1. Consider an alternative model with a parameter restriction imposed of your choosing. Report posterior model probabilities for the unrestricted and restricted models. How did you set the prior model probabilities?

    2. See if you can implement the linear regression model from the previous section using a Metropolis-Hastings sampler instead of a Gibbs Sampler, and then implement the Chib and Jeliaskov (2001) and MHME approach to the marginal likelihood. You should be able to do this with relatively minor modifications to the programs used for **this** exercise. Do you get the same answer for the marginal likelihood that you got in the previous section?