Marginal likelihood calculation for gelfand-dey and Chib Method

Liu, Chun

School of Economics and Management, Tsinghua University, China

October 2010

Marginal Likelihood Calculation for Gelfand-Dey and Chib Method

Abstract

One advantage of Bayesian estimation is its solid theoretical ground on model comparison, which relies heavily upon the accurate calculation of marginal likelihood. The Gelfand-Dey (1994) and Chib (1995) methods are two popular means of calculating marginal likelihood. A trade-off exists between these two methods. The Gelfand-Dey method is simpler and faster to conduct, while Chib method is more accurate, yet intricate. In this paper, we compare the two methods by their ability to identify structural breaks in a reduced form volatility model. Using the Markov Chain Monte Carlo method, we demonstrate that the performance of the two methods is fairly close. Since the Chib method is normally more difficult to implement in many econometric problems, it is safe to choose Gelfand-Dey method when calculating marginal likelihood.

Key words: Model Comparison; Structural Break; Heterogeneous Autoregressive Model

JEL: C11, C52

1 Introduction

One advantage of Bayesian estimation is its solid theoretical ground on model comparison, which relies upon the accurate calculation of marginal likelihood (ML). In academia there are two popular methods for calculating ML; the Gelfand-Dey (1994) and Chib (1995) methods.

Gelfand-Dey (GD) is a general method. The GD method is efficient and utilizes the same routines when calculating the ML for different models. Meanwhile, the Chib method is often thought of as a more accurate method of calculating ML since its routine adjusts to specific model forms. The Chib method requires detailed information about the conditional distributions of all parameters. Unfortunately, it is difficult to attain this detailed information in many cases. Furthermore, the Chib method is more demanding for programming and computation since it is necessary to run simulations for each block of parameters. Hence, it is generally accepted that a trade-off exists between the two methods. The GD method is simpler and faster, while the Chib method is more accurate, yet difficult to implement.

In this paper, we use a Markov Chain Monte Carol simulation to compare the performance of the two methods with a reduced-form volatility model as a benchmark. By investigating the power of the two methods to identify the true models when the underlying data generating processes (DGP) are unstable, we find the GD method performs almost identically to the Chib method. Therefore, we believe it is safe to choose the GD method when calculating ML.

2   Settings

We use the Heterogeneous Autoregressive (HAR) model proposed by Corsi (2009) to model volatility. Corsi shows that the HAR model can account for many of the features of volatility, including long memory. With the observation $Y_T = \{y_1, y_2, \cdots, y_T\}$, HAR model is defined as

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-5,t-1} + \beta_3 y_{t-22,t-1} + \epsilon_t, \ \ \epsilon_t \sim NID(0, \sigma^2). \tag{1}$$

where $y_{t-5,t-1}$ and $y_{t-22,t-1}$ are the average value of y_t in the last 5 or 22 days, respectively. This model postulates three factors that affect y_t: the daily effect y_{t-1}, the weekly effect $y_{t-5,t-1}$ and the monthly effect $y_{t-22,t-1}$.

The DGP instability is captured by the change-point model proposed by Chib (1998). Assume there are $m-1$, $m \in \{1, 2, ...\}$ change points. The density of observation y_t depends on parameter $\theta_k = \{\beta_k, \sigma_k^2\}$, $k = 1, ..., m$, whose value changes at the change points and remains constant otherwise. The state of the system is denoted by $S = \{s_1, s_2, \cdots, s_T\}$ where $s_t = k$ indicates that the observation y_t is from regime k and follows the conditional distribution $f(y_t | I_{t-1}, \theta_k)$. The one-step ahead transition probability matrix for s_t is assumed to be

$$P = \begin{pmatrix} p_{11} & p_{12} & 0 & \cdots & 0 \\ 0 & p_{22} & p_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 0 & p_{m-1,m-1} & p_{m-1,m} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

where $p_{i,j} = \Pr(s_t = j | s_{t-1} = i)$ with $j = i$ or $i+1$, and this is the probability of moving from regime i at time $t-1$ to regime j at time t.

In Bayesian framework, the model comparison is through the Bayes factor (BF), which is defined as $BF_{AB} = p(Y_T|A)/p(Y_T|B)$, where $p(Y_T|A)$ and $p(Y_T|B)$ are marginal likelihoods for model A and B respectively. Kass and Raftery (1995) suggest interpreting the evidence for A as: not worth more than a bare mention for $1 \le BF_{AB} < 3$; positive for $3 \le BF_{AB} < 20$; strong for $20 \le BF_{AB} < 150$; and very strong for $BF_{AB} \ge 150$. Equivalently, based on a log scale, $\log(BF_{AB}) > 0$ is evidence in favor of A versus B.

3   GD Method

Gelfand and Dey (1994) propose a method to calculate ML. They show that for any distribution $g(\theta)$ with the domain contained in the posterior distribution, the reciprocal of ML

$$(p(Y_T))^{-1} = E\left[\frac{g(\theta)}{p(Y_T|\theta)p(\theta)}|Y_T\right] \tag{2}$$

where $p(Y_T|\theta)$ and $p(\theta)$ are the likelihood and the prior for the model, respectively. To calculate, the GD method collects all the samples $\{\theta^j\}_{j=1}^R$ from the posterior distribution, and the equation (2) is estimated as

$$\widehat{p(Y_T)}^{-1} = \frac{1}{R}\sum_{j=1}^R \frac{g(\theta^j)}{p(Y_T|\theta^j)p(\theta^j)}. \tag{3}$$

The prior density $p(\theta^j)$ can be evaluated directly and $p(Y_T|\theta^j)$ is calculated by substituting θ^j into the likelihood function. For the details about the estimation, see Liu and Maheu (2008). According to Gelfand and Dey (1994), if $g(\theta)$ is thin-tailed relative to $p(Y_T|\theta)p(\theta)$, then $g(\theta)/(p(Y_T|\theta)p(\theta))$ is bounded above and the estimator is consistent. Therefore, we follow Geweke (1999) and use a truncated normal distribution $N(\theta^*, \Sigma^*)$ for $g(\theta)$, where θ^* and Σ^* are the posterior sample moments calculated as

$$\theta^* = \frac{1}{R}\sum_{j=1}^R \theta^j, \text{ and } \Sigma^* = \frac{1}{R}\sum_{j=1}^R (\theta^j - \theta^*)(\theta^j - \theta^*)'. \tag{4}$$

When θ^j in the domain of the truncated normal Θ, $g(\theta^j) = c^{-1}\phi(\theta^j)$ where $\phi()$ is the cdf of the truncated normal, and it is 0 otherwise. The constant c ensures $g(\theta)$ integrates to 1.[1] The domain Θ is defined as

$$\Theta = \left\{\theta : (\theta^j - \theta^*)'(\Sigma^*)^{-1}(\theta^j - \theta^*) \leq \chi_\alpha^2(d)\right\} \tag{5}$$

where d is the dimension of the parameter vector and $\chi_\alpha^2(d)$ is the α quantile of the chi-squared distribution with d degrees of freedom. As a smaller α produces a better behaved ratio in equation (2) at a cost of more simulation errors because more draws are dropped, in practice, 0.5, 0.75 and 0.9 are popular selections for α.

4   Chib Method

Chib (1995) advocates a way to calculate the ML. A rearrangement of Bayes rule gives the ML, $p(Y_T)$, as

$$p(Y_T) = \frac{p(Y_T|\theta)p(\theta)}{p(\theta|Y_T)} \tag{6}$$

---

[1]Empirically, c can be obtained by taking a large number of draws from the multivariate normal distribution $N(\theta^*, \Sigma^*)$ and counting the frequency of draws that are in the domain of $g(\theta)$.

where $p(\theta|Y_T)$ is the posterior ordinate. Although this equation is valid for any value θ in the parameter space, a point with high posterior mass will tend to provide a more accurate estimate. Therefore, we select the posterior mean denoted as θ^*. Then

$$\log p(Y_T) = \log p(Y_T|\theta^*) + \log p(\theta^*) - \log p(\theta^*|Y_T). \tag{7}$$

The prior density and the likelihood function can be calculated as above. To calculate the posterior ordinate $p(\theta^*|Y_T)$, note the decomposition

$$p(\theta^*|Y_T) = p(\beta^*|Y_T)\, p(\sigma^{2*}|Y_T, \beta^*)\, p(P^*|Y_T, \beta^*, \sigma^{2*}) \tag{8}$$

where each term on the right hand side can be estimated from Markov Chain Monte Carol simulations. The first term is,

$$p(\beta^*|Y_T) = \int p(\beta^*|Y_T, S, \sigma^2)\, p(S, \sigma^2|Y_T)\, dS d\sigma^2 \tag{9}$$

and can be estimated as

$$\widehat{p(\beta^*|Y_T)} = \frac{1}{R} \sum_{j=1}^{R} p\left(\beta^*|Y_T, \sigma^{2(j)}, S^{(j)}\right) \tag{10}$$

where the draws $\{\sigma^{2(j)}, S^{(j)}\}_{j=1}^{R}$ are available directly from our estimation step.[2] The second term in (8) is equal to

$$p(\sigma^{2*}|Y_T, \beta^*) = \int p(\sigma^{2*}|Y_T, \beta^*, S)\, p(S|, Y_T, \beta^*)\, dS. \tag{11}$$

To obtain the draws from $p(S|Y_T, \beta^*)$, we run an additional reduced Gibbs sampling conditional on β^*; i.e. we run all the sampling again except that we do not draw the values for β but fix them to be β^*. From this reduced run simulation, $\{S^{(j)}\}_{j=1}^{R}$ is used in

$$\widehat{p(\sigma^{2*}|Y_T}, \beta^*) = \frac{1}{R} \sum_{j=1}^{R} p\left(\sigma^{2*}|Y_T, \beta^*, S^{(j)}\right). \tag{12}$$

The last term of (9) can be calculated similarly.

5 Monte Carlo Simulation

We consider the HAR model based on the equation (1) and use two structural break specifications:

- Specification I: all parameters θ change in each regime.

- Specification II: Only β_0 and β_1 change.

---

[2]The conditional density of β and σ^2 is determined by the standard Bayesion estimation equations.

In specification I, the time series properties change greatly and the change points can be identified relatively easily. In specification II, it is challenging to capture the change points when the model structure is more stable. To make our simulation empirically realistic, we select parameter values of the HAR model which are close to those reported in Andersen, Bollerslev and Diebold (2007). They are listed in Table 1. For example, in specification I, if there is 1 break, β_0 changes from -0.1 to -0.4. If there are two breaks, it changes from -0.1 to -0.4 and finally to -0.1.

For each model, we generate $T = 1000$ observations. The true models we consider include cases of no change point, 1 and 2 change points. When there is 1 change point, the position of this change point follows a uniform distribution $U(0.25 \times T, 0.75 \times T)$. When there are 2 change points, the first one follows $U(0.2 \times T, 0.4 \times T)$ and the second follows $U(0.6 \times T, 0.8 \times T)$. This setting allows us to account for the randomness of the change points as well as ensuring sufficient observations in each regime to conduct an estimation.

In the following, the model specifications assumed are known but the number of structural breaks and model parameters are unknown. For each draw from the DGP, we estimate the change-point model assuming $0, 1, 2,$ and 3 structural breaks. We rank the evidence for the number of break points according to the largest ML. The best change point specification has the largest ML, while the second best has the next largest, etc. We draw a new sample of data from the DGP and repeat this until 100 repetitions are completed. Then we report the frequency of repetitions in which each specification is *best* according to ML.

The comparison results of the two methods are summarized in table (2). Panel A is for specification I and Panel B for specification II. The second column indicates the underlying DGP. The $3 - 6$ columns report the frequency over 100 repetitions in which each model is *best* according to ML. For convenience, bold entries in these cells should be 100 if classification is perfect. For example, the first entry tells when underlying DGP is stable (M_0). Out of 100 repetitions, 100 are correctly identified as no change point while 0 are identified incorrectly as 1, 2, or 3 change points according to the Chib method. The next entry in the table repeats this for a DGP with 1 change point (M_1). Here, 100 times the 1 change point is correctly identified. For this "easy" specification, the model ranking results from the Chib and GD methods are identical. The frequencies in their corresponding cells are all 100.

The last three columns report the average log BF over the 100 repetitions, with $\log BF_{i,j}$ denoting the log BF of the model with i change points to the model with j change points. Those numbers are fairly close across the two methods. For example, when underlying DGP is stable (M_0), the log BF of the model with 1 change point to the stable model $\log BF_{1,0}$ is -17.8632 by the Chib method and -18.5999, -18.5440 and -18.5061 by the GD method with $a = 0.5, 0.75$ and 0.9 respectively. This means although the GD and Chib methods calculate different marginal likelihoods, they provide almost identical evidence for the relative strengthen between two models.

Panel B contains the results from the more challenging specification II, where only the first two parameters are allowed to change. The results are very similar to the first case. The frequencies with which the two methods can correctly identify the underlying

DGP are very close: When underlying DGP is stable, 99 out of 100 repetitions are correct by the Chib method, and $100, 98, 98$ by the GD method. When underlying DGP have 1 or 2 changes points, these numbers are 90 or 95 by the Chib method, and $94, 93, 93$ or $97, 96, 96$ by the GD method respectively. Furthermore, their corresponding log BF are roughly the same. Therefore, model comparison and selection using the two methods reports similar results.

In summary, when calculating ML, the Chib and GD methods with difference α prefer the same candidate model. As a result, it is safe to choose GD method under usual conditions (the parameters are not more than 20).

References

[1] Andersen, T. G., T. Bollerslev and F. X. Diebold, 2007. Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility. Review of Economics and Statistics, 89, 701-720.

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

[3] Chib, S., 1998. Estimation and comparison of multiple change point models. Journal of Econometrics, 86, 221-241.

[4] Corsi, F., 2009. A simple long memory model of realized volatility. Journal of Financial Econometrics, 7(2), 174-196.

[5] Gelfand, A. E. and Dey, D. K., 1994. Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society, B 56, 501-514.

[6] Geweke J., 1999. Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication. Econometric Reviews, 18, 1-126.

[7] Kass, R. E. and A. E. Raftery, 1995. Bayes factors and model uncertainty. Journal of the American Statistical Association, 90, 773-795.

[8] Liu, C. and J. M. Maheu, 2008. Are there Structural Breaks in Realized Volatility? Journal of Financial Econometrics, 6(3), 326-360, Summer, 2008.

Table 1: Parameters Values for Simulations

Changed Par.	β_0			β_1			β_2			β_3			σ^2		
Regimes	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Specification I	$-.1$	$-.4$	$-.1$	.1	.4	.1	.4	.15	.4	.1	.4	.1	.2	.5	.2
Specification II	$-.1$	$-.4$	$-.1$	.1	.4	.1	.25	.25	.25	.2	.2	.2	.2	.2	.2

This table lists the hypothetic parameter values for our Monte Carlo simulation. The first column is the changed parameters. The second column lists the index of the regimes. The first row is the Specification index.

Table 2: Method Comparison Using Different Model Specifications

A: Specification I: All parameters are changed

Methods	DGP	Frequency of Models by ML				Ave. Log Bayes Factor		
		0 CP	1 CP	2 CP	3 CP	$\log BF_{1,0}$	$\log BF_{2,1}$	$\log BF_{3,2}$
$Chib$	M_0	**100**	0	0	0	-17.8632	-15.1169	-15.1169
	M_1	0	**100**	0	0	42.5784	-16.0212	-14.8024
	M_2	0	0	**100**	0	41.3661	44.8382	-18.2613
$G-D$	M_0	**100**	0	0	0	-18.5999	-16.3442	-18.4340
($\alpha = 0.5$)	M_1	0	**100**	0	0	42.6793	-17.7100	-16.0211
	M_2	0	0	**100**	0	41.4677	44.9537	-19.3698
$G-D$	M_0	**100**	0	0	0	-18.5440	-16.3452	-18.3874
($\alpha = 0.75$)	M_1	0	**100**	0	0	42.7026	-17.6752	-16.0227
	M_2	0	0	**100**	0	41.4906	44.9676	-19.3760
$G-D$	M_0	**100**	0	0	0	-18.5061	-16.3325	-18.3713
($\alpha = 0.9$)	M_1	0	**100**	0	0	42.7166	-17.6872	-16.0063
	M_2	0	0	**100**	0	41.5044	44.9754	-19.3593

B: Specification II: Only first two parameters are changed

Methods	DGP	Frequency of Models by ML				Ave. Log Bayes Factor		
		0 CP	1 CP	2 CP	3 CP	$\log BF_{1,0}$	$\log BF_{2,1}$	$\log BF_{3,2}$
$Chib$	M_0	**99**	1	0	0	-6.1262	-5.1291	-5.3248
	M_1	0	**90**	8	2	18.4132	-1.1588	-2.2764
	M_2	0	1	**95**	4	11.2407	17.3668	-1.7567
$G-D$	M_0	**100**	0	0	0	-6.4047	-5.4062	-5.9070
($\alpha = 0.5$)	M_1	0	**94**	6	0	18.5095	-1.5342	-2.7379
	M_2	0	0	**97**	3	9.9774	18.8452	-2.4243
$G-D$	M_0	**98**	2	0	0	-6.3441	-5.3698	-5.8625
($\alpha = 0.75$)	M_1	0	**93**	7	0	18.5339	-1.5018	-2.6975
	M_2	0	0	**96**	4	9.9590	18.9105	-2.4013
$G-D$	M_0	**98**	2	0	0	-6.3037	-5.3470	-5.8571
($\alpha = 0.9$)	M_1	0	**93**	7	0	18.5480	-1.4781	-2.6761
	M_2	0	0	**96**	4	9.9541	18.9412	-2.3863

This table reports the comparison results of calculating ML using GD method and Chib method for two model specifications. The second column indicates the underlying DGP. The 3-6 columns display the number of times in the 100 repetitions when that specification had the largest ML. The last three columns report the average log BF between two models.