Communications of the Korean Statistical Society
2009, Vol. 16, No. 3, 479–485

Efficiency and Robustness of Fully Adaptive Simulated Maximum Likelihood Method

Man-Suk Oh[1,a], Dai-Gyoung Kim[b]

[a]Department of Statistics, Ewha Womans University
[b]Department of Applied Mathematics, Hanyang University

Abstract

When a part of data is unobserved the marginal likelihood of parameters given the observed data often involves analytically intractable high dimensional integral and hence it is hard to find the maximum likelihood estimate of the parameters. Simulated maximum likelihood(SML) method which estimates the marginal likelihood via Monte Carlo importance sampling and optimize the estimated marginal likelihood has been used in many applications. A key issue in SML is to find a good proposal density from which Monte Carlo samples are generated. The optimal proposal density is the conditional density of the unobserved data given the parameters and the observed data, and attempts have been given to find a good approximation to the optimal proposal density. Algorithms which adaptively improve the proposal density have been widely used due to its simplicity and efficiency. In this paper, we describe a fully adaptive algorithm which has been used by some practitioners but has not been well recognized in statistical literature, and evaluate its estimation performance and robustness via a simulation study. The simulation study shows a great improvement in the order of magnitudes in the mean squared error, compared to non-adaptive or partially adaptive SML methods. Also, it is shown that the fully adaptive SML is robust in a sense that it is insensitive to the starting points in the optimization routine.

Keywords: Monte Carlo, importance sampling, marginal likelihood, missing.

1. Introduction

In many statistical models a part of data is unobserved (missing), so the complete data consists of observed and unobserved data. Let x denote a vector of observed data and let u denote a vector of unobserved data. Assume that the complete data (x, u) follows a parametric distribution with density $f(x, u; \theta)$, where θ is a vector of unknown parameters.

The observed likelihood function or the marginal likelihood function of θ is given by

$$l(\theta; x) = \int f(x, u; \theta) du \tag{1.1}$$

and the maximum likelihood estimate(MLE) of θ is obtained by maximizing $l(\theta; x)$ with respect to θ. In many practical cases, however, the MLE of θ is not analytically feasible since the integral in (1.1) is analytically intractable and high dimensional.

Due to recent development of high speed computing facilities, Monte Carlo methods have attracted many researchers and practitioners as a solution to the above integration problem. In particular, simulated maximum likelihood(SML) method estimates the integral by an average based on simulated data

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2007-531-C00016).

[1] Corresponding author: Professor, Department of Statistics, Ewha Womans University, Dae-Hyun Dong 11-1, Seoul 120-750, Korea. E-mail: msoh@ewha.ac.kr

set and approximates the MLE by maximizing the estimated likelihood function. Due to its simplicity and efficiency, SML has been used in many applications (Crepon and Duguet, 1997; Danielsson, 1994; Hurn *et al.*, 2003; Jank and Booth, 2003; Kao *et al.*, 2001; Lee, 1995; Munkin and Trivedi, 1999).

In SML, the integral is often estimated using Importance Sampling(IS) Monte Carlo method. Importance Sampling estimates the marginal likelihood by

$$\hat{l}(\theta; x) = \frac{1}{N} \sum_{i=1}^{N} \frac{f(x, u_i; \theta)}{g(u_i)}, \tag{1.2}$$

where $g(u)$ is a density function of u and $u_1, \ldots, u_N$ is a random sample from a distribution with density $g(u)$. Since u_i follows the density function $g(u)$,

$$E\left[\hat{l}(\theta; x)\right] = \int \frac{f(x, u; \theta)}{g(u)} g(u) du = \int f(x, u; \theta) du = l(\theta; x).$$

Thus, with a large number of simulation samples, $\hat{l}(\theta; x)$ converges to $l(\theta; x)$ and the maximum point of $\hat{l}(\theta; x)$ would be an approximate MLE of θ.

Compared with Markov chain Monte Carlo method, Importance Sampling has several advantages. First, it directly generates random samples of vector u instead of generating elements of u from conditional distributions, so it is often very efficient in high dimensional integrals. Second, it is less restrictive since it does not generate samples from exact (conditional) distributions. Finally, estimation error can be easily computed.

The efficiency of IS depends on the choice of the density function $g(u)$, called the importance sampling density or the proposal density. In SML, the optimal proposal density is a density function which is proportional to $f(x, u; \theta)$ for each given x and θ, *i.e.*, the optimal $g(u)$ is $g^*(u) = f(u|x, \theta)$, where $f(u|x, \theta)$ is a conditional density of u given x and θ.

At $\theta = \theta_{MLE}$, the optimal proposal density is $f(u|x, \theta_{MLE})$ and the estimation error in $\hat{l}(\theta_{MLE}; x)$ is zero, where θ_{MLE} is the MLE of θ (Zhang, 1996; McCulloch, 1997). For this reason, $g^M(u) = f(u|x, \theta_{MLE})$ is called the optimal proposal density. However, since it is optimal only at θ_{MLE}, $g^M(u)$ is only a *locally* optimal proposal density. The *locally* optimal proposal density $g^M(u)$ is not possible to obtain since we don't know θ_{MLE} and/or it is hard to find the closed form of $f(u|x, \theta_{MLE})$. A remedy to this problem is to find an approximation or a guess $\hat{\theta}$ of θ_{MLE} and use a Laplace approximation of $f(u|x, \hat{\theta})$,

$$f_{Lap}\left(u|x, \hat{\theta}\right) = \text{density function of } N\left(\mu\left(\hat{\theta}\right), \Sigma\left(\hat{\theta}\right)\right),$$

where $\mu(\hat{\theta})$ and $\Sigma(\hat{\theta})$ are the Laplace approximations to the mean and variance of $f(u|x, \hat{\theta})$, respectively.

However, Jank (2006) showed that the accuracy of SML highly depends on the choice of $\hat{\theta}$ and that SML may lead to significantly misleading results if $\hat{\theta}$ is far from θ_{MLE}. This implies that in the beginning we need to have good information about θ_{MLE}, which is generally unavailable. To overcome this problem of lacking information about θ_{MLE}, adaptive Importance Sampling schemes have been proposed. Instead of using one proposal density $f_{Lap}(u|x, \hat{\theta})$ throughout the simulation, Jank (2006) divides the simulation into several stages and at each stage it updates $\hat{\theta}$ by using an estimate of θ_{MLE} from the previous stage, to obtain a better approximation to the locally optimal proposal density.

The adaptive SML significantly improves the efficiency of SML. However, it has some key issues. First, numerical optimization schemes in SML require computation of $\hat{l}(\theta; x)$ not only at $\hat{\theta}$ or θ_{MLE} but at various other values of θ. However, the approximate locally optimal proposal density $f_{Lap}(u|x, \hat{\theta})$

is optimal only for $\hat{l}(\hat{\theta}; x)$, not for general $\hat{l}(\theta; x)$. Even for θ close to $\hat{\theta}$, $l(\hat{\theta}; x)$ and $l(\theta; x)$ may be quite different and $f_{Lap}(u|x, \hat{\theta})$ may not be a good proposal density for estimating $l(\theta; x)$, $\theta \neq \hat{\theta}$. In other words, Jank's adaptive scheme tries to find an approximation to $f(u|x, \theta_{MLE})$ but $f(u|x, \theta_{MLE})$ is optimal only for $\hat{l}(\theta_{MLE}; x)$ and it may not be a good choice for $\hat{l}(\theta; x)$, $\theta \neq \theta_{MLE}$, which is repeatedly computed in SML. Second, it uses $\hat{\theta}$ obtained from the previous stage and restart SML in the current stage, thus the final estimate of θ_{MLE} is obtained using random samples only from the last stage, wasting precious random samples from all the previous stages.

A fully adaptive SML attempts to find a good approximation to $f(u|x, \theta)$ for each θ, rather than focusing on θ_{MLE}. In other words, for each θ, $g(u|\theta) = f_{Lap}(u|x, \theta)$ is used as a proposal density in Importance Sampling. Note that the proposal density of u depends on θ, Equation (1.2) is then replaced by

$$\hat{l}(\theta; x) = \frac{1}{N} \sum_{i=1}^{N} \frac{l(\theta; x, u_i)}{g(u_i|\theta)}.$$

For each θ, $\hat{l}(\theta; x)$ converges to $l(\theta; x)$ and the variance of the estimate would be zero if the Laplace approximation is exact.

In stead of using one proposal density throughout the simulation or in each stage of adaptive SML, the fully adaptive scheme uses a new proposal density $g(u|\theta)$ which is an approximation to the optimal proposal density $f(u|x, \theta)$ for each and every new value of θ. In other words, it is fully adaptive in a sense that it updates g for each new value of θ. Thus, compared with Jank's adaptive SML which we call a partially adaptive SML, in the fully adaptive SML each stage consists of one iteration and it updates g instead of $\hat{\theta}$ so that g is approximately optimal for computing $l(\theta; x)$ for each given θ. Thus, it highly improves the accuracy of Importance Sampling estimate $\hat{l}(\theta; x)$ for each θ and hence improves the efficiency of SML, as will be shown in a simulation study given in Section 3. Also, the fully adaptive SML uses a new parameter values of g at each iteration but uses all the random samples u_i in the estimation of the marginal likelihood. So it does not waste random samples and it does not require appropriate sample sizes in stages.

Though non-adaptive SML is considered as the main SML algorithm in the statistical literature, the fully adaptive SML has been used in some applications such as state space model, multinomial probit model, and econometrics literature (Durbin and Koopman, 1997, 2000; Stern, 1997; Richard and Zhang, 2007). However, the fully adaptive SML has not been well recognized in statistical literature and the efficiency of fully adaptive SML has not been well understood. Only recently Brinch (2008) described the fully adaptive SML in detail, which he named tilted importance sampling, and developed the distinction between the fully adaptive SML and the simple SML.

In this paper, we describe the fully adaptive SML in a way to help understanding why it is more efficient than the simple non-adaptive SML or partially adaptive SML, and study its estimation performance compared with the non-adaptive and partially adaptive SML via a simulation study. We also study its robustness to the initial values of the parameters of interest in the estimation process.

This paper is organized as follows, In Section 2, details of the fully adaptive SML method is described. In Section 3, the efficiency and robustness comparison between the fully adaptive SML, the non-adaptive and partially adaptive SML is given via a simulation study. A brief summary and discussion is given in Section 4.

2. Fully Adaptive SML

Laplace approximation to $f(u|x,\theta)$ is a density function of Normal distribution with mean and variance equal to the mode and curvature at the mode of $f(u|x,\theta)$, respectively. Let $L(\theta;x,u) = \log f(x,u;\theta)$ then the Laplace approximation of the mean $\mu(\theta)$ is $\hat{u}(\theta)$ where $\hat{u}(\theta)$ is the maximizer of $L(\theta;x,u)$ with respect to u, satisfying $\partial/\partial u L(\theta;x,u)|_{u=\hat{u}(\theta)} = 0$. The Laplace approximation of the variance $\Sigma(\theta)$ is minus the inverse Hessian of L at $u = \hat{u}(\theta)$. Note that the mean and variance of u depend on θ so that $g(u|\theta)$ has the optimal mean and variance for each given θ.

SML maximizes $\hat{l}(\theta;x)$ where $\hat{l}(\theta;x)$ is given in (1.2) with $u_i \sim N(\mu(\theta),\Sigma(\theta))$. However, in optimizing Monte Carlo estimate $\hat{l}(\theta;x)$ with respect to θ, $\hat{l}(\theta;x)$ should depend only on θ for stability in the optimization routine. In this case, $\hat{l}(\theta;x)$ depends not only on θ but also on the random samples u_i, resulting in a different value of $\hat{l}(\theta;x)$ even for the same θ and N and hence instability in the optimization. To avoid the instability, the random samples used in the Monte Carlo estimation should be fixed for all iterations in the optimization. This can be done as follows: generate a set of random variables $\{z_i\}_{i=1}^N$ from $N(0,I)$ and keep using it in each step of optimization by using the transformation

$$u_i = T(\theta)z_i + \mu(\theta),$$

where $T(\theta)$ is a lower triangular matrix such that $T(\theta)T'(\theta) = \Sigma(\theta)$. In this way, a fixed set of $\{z_i\}$ is used in each iteration, yet u_i follows $N(\mu(\theta),\Sigma(\theta))$ and the SML essentially optimizes a fixed (non-random) function $\hat{l}(\theta;x)$ of θ. Moreover, use of the same set of random variates may reduce the cost of generating random samples in each iteration and hence increase the efficiency of SML.

Now the fully adaptive SML algorithm is summarized.

- Step 1: Generate z_i from $N(0,I)$, for $i = 1,\ldots,N$ and store them in a table.
- Step 2: Maximize

$$\hat{l}(\theta;x) = \frac{1}{N}\sum \frac{l(\theta;x,u_i)}{g(u_i|\theta)}$$

over θ, where $g(u|\theta) = f_{Lap}(u|x,\theta)$ = density function of $N(\mu(\theta),\ \Sigma(\theta))$, $u_i = T(\theta)z_i + \mu(\theta)$ and $T(\theta)$ is a lower triangular matrix such that $T(\theta)T'(\theta) = \Sigma(\theta)$.

In Importance Sampling, tails of the proposal density g should be heavier than or equal to those of $f(u|x,\theta)$ to avoid variance inflation in Monte Carlo estimates (Oh and Berger, 1992). Normal density converges to 0 fast for extreme values in either direction and may cause the variance inflation problem. A solution to this problem is to use a t density with a small degrees of freedom in place of normal distribution in the Laplace approximation of $f(u|x,\theta)$.

3. Performance Evaluation and Robustness via a Simulation Study

Suppose that the binary responses x_{ij}, $i = 1,\ldots,q$, $j = 1,\ldots,n$, are obtained from the following logistic-normal model: X_{ij} are conditionally independent with $X_{ij}|u_i \sim$ Bernoulli(π_{ij}), where

$$\text{logit}(\pi_{ij}) = \log\frac{\pi_{ij}}{1-\pi_{ij}} = \beta t_{ij} + u_i, \quad u_i \sim N(0,\sigma^2).$$

Jank (2006) generated data from the logistic-normal model with $q = 10$, $n = 15$, $\beta = 5$, $\sigma^2 = 0.5$, $t_{ij} = j/15$ and the date are presented in Table 1.

Table 1: Simulated logistic-normal data x_{ij}

								j							
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	0	0	0	1	1	0	1	1	1	1	1	1	1
2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1
6	0	0	0	1	0	1	1	1	0	1	1	1	1	1	1
7	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

In this model the integral involved in the marginal likelihood is low dimensional and one can obtain a very accurate estimate of MLE of the model parameters $\theta = (\beta, \sigma^2)$. The (almost) exact MLE of θ is $\theta_{MLE} = (\beta_{MLE}, \sigma^2_{MLE}) = (6.132, 1.766)$.

Assuming that we don't know the MLE, we applied the non-adaptive SML, the partially adaptive SML of Jank (2006), and the fully adaptive SML with $N = 600$ and 6000. In the partial adaptive SML, we tried 3 stages with 100, 200, 300 sample sizes in the three stages, respectively, when $N = 600$, and 1000, 2000, 3000 sample sizes when $N = 6000$.

To study robustness of the schemes, we used four different starting points of $\theta = (\beta, \sigma^2)$ as shown in Tables 2 and 3, and used DUMPOL, the optimization routine for non-smooth functions in IMSL (1989) library, in all the three SML methods. We repeated the simulation 100 times and averaged the Mean Squared Error(MSE), $\text{MLE}_\beta = (\hat{\beta} - \beta_{\text{MLE}})^2$ and $\text{MLE}_{\sigma^2} = (\hat{\sigma^2} - \sigma^2_{MLE})^2$, and the computing time T in seconds, which are reported in Table 2. All the computations are done using PC with Pentium IV Processor. Considering both MSE and computing time T, we also calculated an efficiency measure

$$Eff = \frac{1}{\text{MLE} \times T}$$

for β and σ^2, and they are presented in Table 3.

The results clearly show great improvements of the MSE and the efficiency in the fully adaptive SML method. Since the fully adaptive SML computes $\mu(\theta)$ and $\Sigma(\theta)$ for each given θ, the computing time of fully adaptive method is larger than those of other methods. However, it reduces MSE in the order of magnitudes and hence greatly improved the overall efficiency.

In the non-adaptive scheme and the partially adaptive scheme, the MSE's do not always get smaller as the Monte Carlo sample size N gets larger. Even when there is a reduction in MSE's for the larger N the reduction is very small compared to the increased computing time, resulting in the decrease of efficiency. Note that the efficiency is smaller when $N = 6000$ in all the cases shown in Table 3. However, in the fully adaptive scheme the MSE's are much smaller for the larger N, resulting in the increase of efficiency, in all the cases. Moreover, the MSE's in the fully adaptive scheme with $N = 600$ are much smaller than those in the non- or partially adaptive scheme with $N = 6000$. These imply that simply increasing the Monte Carlo sample size does not necessarily improve the accuracy of SML and an appropriate scheme for finding a good proposal density is much more important.

Note also that the fully adaptive SML is insensitive to the initial values of the parameters. This may be because the fully adaptive scheme computes a very accurate estimate of the marginal likelihood, the function to be optimized, and provides a good direction to the numerical optimization routine.

We also tried t density with various degrees of freedom in these examples but the results are about the same, implying that the normal proposal density seems to be a reasonable choice.

Table 2: MSE and computing time

initial	N	Non-adaptive			Partially Adaptive			Fully-adaptive		
		MSE_β	MSE_{σ^2}	T	MSE_β	MSE_{σ^2}	T	MSE_β	MSE_{σ^2}	T
(1, 0)	600	1.7072	1.9372	4.1	0.2994	0.5037	3.4	0.00001	0.00068	11.6
	6000	1.0887	1.4413	27.6	0.0873	0.2180	24.3	0.00001	0.00008	64.5
(2, 1.2)	600	0.3403	0.4784	3.1	0.0351	0.2033	3.0	0.00019	0.00135	11.6
	6000	0.1600	0.2315	30.4	0.0186	0.5889	28.3	0.00001	0.00008	77.7
(4, 1.4)	600	0.1383	0.4010	3.3	0.1536	0.6025	3.0	0.00007	0.00059	14.8
	6000	0.1280	0.2028	30.3	0.0195	0.2982	28.1	0.00001	0.00007	72.6
(6.132, 1.766)	600	0.0125	0.0306	3.0	0.0386	0.2764	3.0	0.00014	0.00118	9.52
	6000	0.0061	0.1483	29.2	0.0136	0.2451	27.4	0.00001	0.00007	72.4

Table 3: Efficiency

initial	N	Non-adaptive		Partially Adaptive		Fully-adaptive	
		Eff_β	Eff_{σ^2}	Eff_β	Eff_{σ^2}	Eff_β	Eff_{σ^2}
(0, 1)	600	0.1425	0.1256	0.9880	0.5870	926.5	124.8
	6000	0.0332	0.0251	0.4720	0.1890	1408.9	186.3
(2, 1.2)	600	0.9509	0.6765	9.5652	1.6510	476.4	66.3
	6000	0.2056	0.1419	1.8966	0.0600	1195.2	156.4
(4, 1.4)	600	2.2178	0.7649	2.1775	0.5551	901.8	113.9
	6000	0.2585	0.1631	1.8215	0.1194	1410.1	195.4
(6.132, 1.766)	600	26.7940	10.9410	8.5216	1.1902	725.2	89.3
	6000	5.6080	0.2308	2.6852	0.1488	1512.6	197.2

4. Summary and Discussion

Simulated Maximum Likelihood estimates an intractable marginal likelihood via Importance Sampling Monte Carlo method and optimizes the Monte Carlo estimate of the likelihood to find the maximum likelihood estimate of parameters of interest θ, when data contains missing observations. A key issue in Importance Sampling is an appropriate choice of the proposal density function from which simulation samples are generated. Efforts have been given to find a good approximation to the locally optimal proposal density which is optimal at the MLE itself. However, the locally optimal proposal density is optimal only at the MLE while optimization routines require computation of the marginal likelihood at many other values of θ.

In this paper, a fully adaptive SML method is described in a way to help understanding its distinction from the non- or partially adaptive SML and its source of improved efficiency. It provides a good approximation to the optimal proposal density at each given θ. Thus, while the optimization routines explore the space of θ, the fully adaptive SML provides a good estimate of the function to be optimized at each given θ. In this sense, the proposal density is globally optimal rather than locally optimal.

Also, to stabilize the optimization of Monte Carlo estimates, it fixes standard normal random samples in Monte Carlo estimation but uses transformations appropriate for each given θ.

The fully adaptive SML is very simple and a simulation study shows a dramatic improvement in the accuracy and efficiency compared to the currently available non-adaptive and a partially adaptive SML methods. Moreover, it contains almost the same efficiency even for very bad initial values of θ, implying robustness of the method.

In this paper, we use a normal or t density function for the functional form of the proposal density. Other density functions such as gamma or beta may be used if we have some information about the shape of the conditional density of u given x and θ.

References

Brinch, C. N. (2008). Simulated maximum likelihood using tilted importance sampling, *Statistics Norway, Research Department, Discussion papers No. 540.*

Crepon, B. and Duguet, E. (1997). Research and development, competition and innovation pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity, *Journal of Econometrics,* **79**, 355–378.

Danielsson, J. (1994). Stochastic volatility in asset prices estimation with simulated maximum likelihood, *Journal of Econometrics,* **64**, 375–400.

Durbin, J. and Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models, *Biometrika,* **84**, 669–684.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives, *Journal of the Royal Statistical Society, Series B,* **62**, 3–56.

Hurn, A. S., Lindsay, K. A. and Martin, V. L. (2003). On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations, *Journal of Time Series Analysis,* **24**, 45–63.

IMSL (1989). *User's Manual,* IMSL, Houston, Texas.

Jank, W. (2006). Efficient simulated maximum likelihood with an application to online retailing, *Statistics and Computing,* **16**, 111–124.

Jank, W. and Booth, J. (2003). Efficiency of Monte Carlo EM and simulated maximum likelihood in two-stage hierarchical models, *Journal of Computational and Graphical Statistics,* **12**, 214–229.

Kao, C., Lee, L. F. and Pitt, M. M. (2001). Simulated maximum likelihood estimation of the linear expenditure system with binding non-negativity constraints, *Annals of Economics and Finance,* **2**, 203–223.

Lee, L. F. (1995). Asymptotic bias in simulated maximum likelihood estimation of discrete choice models, *Econometric Theory,* **11**, 437–483.

Lee, L. F. (1997). Simulated maximum likelihood estimation of dynamic discrete choice statistical models some Monte Carlo results, *Journal of Econometrics,* **82**, 1–35.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association,* **92**, 162–170.

Munkin, M. K. and Trivedi, P. K. (1999). Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application, *Econometrics Journal,* **2**, 29–48.

Oh, M. S. and Berger, J. O. (1992). Adaptive importance sampling in monte carlo integration, *Journal of Statistical Computation and Simulation,* **41**, 143–168.

Richard, J. F. and Zhang, W. (2007). Efficient high-dimensional importance sampling, *Journal of Econometrics,* **141**, 1385–1411.

Stern, S. (1997). Simulation-based estimation, *Journal of Econometric Literature,* **35**, 2006–2039.

Zhang, P. (1996). Nonparametric importance sampling, *Journal of the American Statistical Association,* **91**, 1245–1253.

Received February 2009; Accepted April 2009