Exponential tilting in approximate posterior computation

Samer A Kharroubi*
Department of Nutrition and Food Sciences,
Faculty of Agricultural and Food Sciences,
American University of Beirut.

January 9, 2017

Abstract

We explore the use of importance sampling based on signed root log-likelihood ratios incorporating exponential tilting to obtain versions of various asymptotic formulae for Bayesian computation. The modifications are designed to avoid repeated evaluation of conditional maxima of the likelihood function. This leads to a more stable computational procedure as well as significantly reducing computational time. A substantial reduction in sampling variability can also be achieved by incorporating antithetic variates. Implementation of the modified signed root based importance sampler and antithetic variates are illustrated by a censored regression model.

Keywords: Approximate Bayesian inference; Higher-order asymptotics; Exponential tilting; Signed root log-likelihood ratio; Importance sampling; Variance reduction.

1 Introduction

Accurate asymptotic formulae for posterior expectations and predictive distributions were obtained in Sweeting (1996), Sweeting and Kharroubi (2003) and Kharroubi and Sweeting (2010). One difficulty with applying these formulae is the need for repeated computation of conditional maxima of the likelihood function in multiparameter cases. This can be particularly problematic when it is required to invert the signed root log-likelihood ratio, since a maximization procedure has to be included within a nonlinear inversion routine, as pointed out in Sweeting (1996) and the ensuing discussion. Recently, Kharroubi and Sweeting (2016) used exponential tilting to develop alternative asymptotic approximations for posterior expectations, predictive distributions and marginal posterior distributions that do not require any conditional maximization. The new tilted approximations are shown to be particularly relevant for the signed root based importance

*Correspondence to: Dr Samer A Kharroubi, Department of Nutrition and Food Sciences, Faculty of Agricultural and Food Sciences, American University of Beirut, P.O.BOX: 11-0236, Riad El Solh 1107-2020 Beirut, Lebanon. Phone: +961 (1) 350000, Ext: 4541. Email: sk157@aub.edu.lb.

sampler and antithetic variates of Kharroubi and Sweeting (2010). A number of researchers have obtained useful asymptotic approximations in statistics based on exponential tilting; see for example, Schennach (2005, 2007), Cerquetti (2007) and Sweeting and Kharroubi (2005).

In the present paper we follow on from Kharroubi and Sweeting (2016) work by using exponential tilting to develop a modified signed root based importance sampler for moderately parameterized models that give rise to exact (that is, simulation-consistent) computation of various posterior quantities of interest, including posterior moments and marginal posterior densities. It is further shown that by incorporating antithetic variates we can achieve a substantial reduction in sampling variability with relatively short runs of the schemes. The potential of the methodology is explored throughout the paper using a censored regression model.

2 Signed roots using exponential tilting

Suppose that the data x arise from a parametric model with $\theta = (\theta^1, \ldots, \theta^d) \in \Omega \subset \mathbb{R}^d$, where $d \geq 1$, and that the associated likelihood function $L(\theta)$ is available. Suppose the availability of a prior density $\lambda(\theta)$ of θ, continuous and positive throughout Ω. Then the posterior density of θ is

$$p(\theta|x) = c^{-1} L(\theta)\lambda(\theta)\,, \tag{1}$$

where $c = \int L(\theta)\lambda(\theta)\, d\theta$. Assume that the likelihood function is twice continuously differentiable. Let $l(\theta) = \log L(\theta)$ and define $\dot{l}(\theta) = dl(\theta)/d\theta = (l_1(\theta), \ldots, l_d(\theta))^T$, where $l_j(\theta) \equiv \partial l(\theta)/\partial\theta^j$, $j = 1, \ldots, d,$. Further define $\ddot{l}(\theta) = d^2 l(\theta)/d\theta^2$, $j(\theta) = -\ddot{l}(\theta)$ and $J = j(\hat{\theta})$, the observed information. As in Sweeting and Kharroubi (2003), for $1 \leq i \leq d$ we let $\theta_i = (\theta^1, \ldots, \theta^i)$ be the vector of the first i components of θ and $\theta^{(i)} = (\theta^i, \ldots, \theta^d)$ the vector of the last $d - i + 1$ components.

We assume that, for each $1 < i \leq d$ and fixed θ_{i-1}, there exists a local conditional maximizer $\hat{\theta}^{(i)}(\theta_{i-1}) = (\hat{\theta}^i(\theta_{i-1}) \ldots, \hat{\theta}^d(\theta_{i-1}))$ of $L(\theta)$. For $1 \leq j \leq d$, $\hat{\theta}^j(\theta_{i-1})$ will denote the jth component of $(\theta_{i-1}, \hat{\theta}^{(i)}(\theta_{i-1}))$; thus $\hat{\theta}^j(\theta_{i-1}) = \theta^j$ for $1 \leq j < i$.

It follows from Kharroubi and Sweeting (2016) that the first-order term in the Taylor expansion of the jth component of the conditional maximizer of $l(\theta)$ given θ_i about $\theta_i = \hat{\theta}_i$ is

$$\bar{\theta}^j(\theta_i) = \hat{\theta}^j + \sum_{k=1}^{i} \hat{\theta}^j_{ik}(\theta^k - \hat{\theta}^k) \tag{2}$$

for $1 \leq i < j \leq d$, where $\hat{\theta}^j_{ik} = \partial\hat{\theta}^j(\theta_i)/\partial\theta^k$ evaluated at $\theta_i = \hat{\theta}_i$. Further write $\bar{\theta}^{(j)}(\theta_i) = (\bar{\theta}^j(\theta_i), \ldots, \bar{\theta}^d(\theta_i))$ for $1 \leq i < j \leq d$.

As in Kharroubi and Sweeting (2016), we will use the following convention: for any function $g(\theta)$ and $1 \leq i < d$, $g(\theta_i)$ will denote $g(\theta_i, \bar{\theta}^{(i+1)}(\theta_i))$, where $g(\theta_0)$ is understood to be $g(\hat{\theta})$. That is, we simply replace $\hat{\theta}^{(i+1)}(\theta_i)$ by $\bar{\theta}^{(i+1)}(\theta_i)$. Now, define the functions

$$h^i(\theta) = \sum_{j=1}^{d} c^j_i l_j(\theta) = c_i^T \dot{l}(\theta)\,, 1 \leq i \leq d\,,$$

where $c_i = (c_i^1, \ldots, c_i^d)^{\mathrm{T}}$ and $c_i^j = 0\,(j < i)$, $c_i^i = 1$, $c_i^j = \hat{\theta}_{ii}^j\,(j > i)$. The exponentially tilted log-likelihood function is defined by

$$\bar{l}(\theta) = l(\theta) - \sum_{j=1}^{d} h^j(\theta_{j-1})\{\theta^j - \bar{\theta}^j(\theta_{j-1})\} \tag{3}$$

on setting $\theta_0 = \hat{\theta}$. Note that $\hat{\theta}$ is also the maximizer of $\bar{l}(\theta)$ since, for $1 \le j \le d$, $\bar{\theta}^j(\hat{\theta}_{j-1}) = \hat{\theta}^j$. It is shown in Kharroubi and Sweeting (2016) that, for $1 < i \le d$, $\bar{\theta}^{(i)}(\theta_{i-1})$ is the unique local conditional maximizer of $\bar{l}(\theta)$ given θ_{i-1}.

Now define the functions $H^i(\theta_i) = \exp\{h^i(\theta_{i-1})\{\theta^i - \bar{\theta}^i(\theta_{i-1})\}$ for $1 \le i \le d$ and let $H(\theta) = \prod_{i=1}^d H^i(\theta_i)$. Then from (3) the exponentially tilted likelihood function is given by $\bar{L}(\theta) = L(\theta)/H(\theta)$, so the exponentially tilted prior density $\bar{\lambda}(\theta) = \lambda(\theta)H(\theta)$.

Specifically, for $1 \le i \le d$ define the tilted log-likelihood ratios

$$\bar{w}^i(\theta_i) = 2\{\bar{l}(\theta_{i-1}) - \bar{l}(\theta_i)\} = 2\{l(\theta_{i-1}) - l(\theta_i) + h^i(\theta_{i-1})(\theta^i - \bar{\theta}^i(\theta_{i-1}))\}$$

and the tilted signed root log-likelihood ratios $\bar{r}(\theta) = (\bar{r}^1(\theta_1), \ldots, \bar{r}^d(\theta_d))$ by

$$\bar{r}^i(\theta_i) = \mathrm{sign}\{\theta^i - \bar{\theta}^i(\theta_{i-1})\}\{\bar{w}^i(\theta_i)\}^{1/2}$$

3 Importance sampling based on tilted signed roots

Consider the posterior expectation $\mu = E\{v(\theta)|x\}$ of the smooth function $v(\theta)$. Let $g(\theta)$ be any density function from which it is easy to sample. It follows that, since

$$\mu = c^{-1} E\left\{ v(\theta) \frac{L(\theta)\lambda(\theta)}{g(\theta)} \right\},$$

the constant of proportionality c in (1) and μ are consistently estimated by, respectively

$$\hat{c} = (km)^{-1} \sum_{j=1}^{m} u_j \tag{4}$$

$$\hat{\mu} = \sum_{j=1}^{m} v(\theta_{[j]}) w_j \tag{5}$$

where $\theta_{[1]}, \ldots, \theta_{[m]}$ are m independent draws from $g(\theta)$, $u_j = L(\theta_{[j]})\lambda(\theta_{[j]})/h(\theta_{[j]})$ are the importance weights, $w_j = u_j / \sum_{i=1}^m u_i$ the normalized importance weights and $g(\theta) = kh(\theta)$.

The importance sampling scheme of Kharroubi and Sweeting (2010) entails a large amount of conditional maximization so the use of exponential tilting is highly relevant here. We therefore modify this scheme by replacing $r(\theta)$ with $\bar{r}(\theta)$. Thus, for $1 \le i \le d$, $\theta^i \equiv \theta^i(\bar{R}_i)$ are defined by inversion of $\bar{r}^i(\theta_i) = \bar{R}^i$ for fixed θ_{i-1}, where the $\bar{R}^i$ are independently sampled from the standard normal distribution. The simulation-consistent estimators of c and μ are then given by (4) and (5) respectively with $k = (2\pi)^{d/2}\bar{L}(\hat{\theta})$ and $u_j = u(\theta_{[j]})$, where

$$u(\theta) = \bar{\lambda}(\theta) \prod_{i=1}^{d} \left\{ \frac{\bar{r}^i(\theta_i)}{-\bar{l}_i(\theta_i)} \right\}.$$

Finally, it follows as in Kharroubi and Sweeting (2010) that a simulation-consistent estimator of the marginal posterior density $p(\theta_i|x)$ of the first i components of θ is given by

$$\bar{p}(\theta_i|x) = \frac{(2\pi)^{(d-i)/2}}{\hat{c}m} \sum_{j=1}^{m} \frac{\bar{L}(\theta_{[j]i})\bar{L}(\theta_i, \tilde{\theta}_{[j]}^{(i+1)})\bar{\lambda}(\theta_i, \tilde{\theta}_{[j]}^{(i+1)})}{\bar{L}(\theta_{[j]})} \prod_{k=i+1}^{d} \frac{-\bar{R}_{[j]}^{k}}{\bar{l}_k(\theta_{[j]k})}\,, \tag{6}$$

where, for $j = 1, \ldots, m$, $\tilde{\theta}_{[j]}^{(i+1)} = \theta_{[j]}^{(i+1)} + \bar{\theta}^{(i+1)}(\theta_i) - \bar{\theta}^{(i+1)}(\theta_{[j]i})$.

It is also possible to employ antithetic variate techniques for variance reduction, simply by replacing $r(\theta)$ in Kharroubi and Sweeting (2010) by $\bar{r}(\theta)$. This can achieve high accuracy with a much smaller importance sample. Variance reduction techniques in conjunction with importance sampling can be found in Hammersley and Handscomb (1964) and Evans and Swartz (2000).

Here there is a natural antithetic variate, namely $\tilde{\bar{R}} = -\bar{R}$, which also has the multivariate standard normal distribution. Suppose that $\theta_{[j]} = \theta(\bar{R}_{[j]})$ and $\tilde{\theta}_{[j]} = \theta(\tilde{\bar{R}}_{[j]})$ for $j = 1, \ldots, m$. Then, from (4) and (5), the Monte Carlo estimators of the normalizing constant and posterior expectation under antithetic importance sampling are

$$\hat{c} = (km)^{-1} \sum_{j=1}^{m} \frac{1}{2}(u_j + \tilde{u}_j) \tag{7}$$

$$\hat{\mu} = \sum_{j=1}^{m} \left\{ v(\theta_{[j]}) w_j + v(\tilde{\theta}_{[j]}) \tilde{w}_j \right\} \tag{8}$$

respectively, where $\tilde{u}_j = L(\tilde{\theta}_{[j]})\lambda(\tilde{\theta}_{[j]})/g(\tilde{\theta}_{[j]})$ are the importance weights, $w_j = u_j / \sum_{i=1}^{m}(u_i + \tilde{u}_i)$ and $\tilde{w}_j = \tilde{u}_j / \sum_{i=1}^{m}(u_i + \tilde{u}_i)$ are the normalized importance weights.

EXAMPLE. *Normal regression model with censored data*

We consider the censored regression model for the failure data given by Crawford (1970). These data arise from temperature accelerated life tests on electrical insulation in 40 motorettes. Ten motorettes were tested at each of four temperatures in degrees Centigrade (150°, 170°, 190° and 220°), resulting in a total of 17 failed units and 23 unfailed (i.e. censored) units.

As in Schmee and Hahn (1979), the model for the data takes the form

$$X_i = \beta_0 + \beta_1 v_i + \sigma\epsilon_i, \quad i = 1, \ldots, 40\,,$$

where X_i is the $\log_{10}(i\text{th failure time})$, with time in hours, $v_i = 1000/(\text{temperature} + 273.2)$ and the errors ϵ_i are independent standard normal. Reordering the data so that the first k observations are uncensored, with observed log-failure times X_i, and the remaining $n - k$ are censored at times Z_i, the log-likelihood function is

$$-k\log\sigma - \frac{1}{2}\sum_{i=1}^{k}\left(\frac{X_i - \beta_0 - \beta_1 v_i}{\sigma}\right)^2 + \sum_{i=k+1}^{n} \log\left\{1 - \Phi\left(\frac{Z_i - \beta_0 - \beta_1 v_i}{\sigma}\right)\right\}\,,$$

For computational convenience we work with the parameterisation $\theta = (\beta_0, \beta_1, \phi)$, where $\phi = \log\sigma$. Here we find that $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\phi}) = (-6.0193, 4.3112, -1.3502)$.

For the purpose of illustration, we examine the accuracy of (5) with $\lambda(\theta) \propto 1$ and $v(\theta) = \beta_0 + \beta_1 + \exp(\phi)$. The signed root based simulation algorithm using exponential tilting was run with $m = 1000$ and the posterior expectation of $v(\theta)$ was found to be -1.4980. The corresponding expectation obtained from the non-tilted algorithm was -1.4995. The exact posterior expectation of $v(\theta)$, obtained via the data augmentation scheme of Tanner and Wong (1987), is -1.4986. Even with m as low as 1000 both schemes produce very good approximations in this case. However, generating a random sample of size $m = 1000$ via the tilted scheme achieves about a 60% time reduction compared with the untilted scheme. Another major advantage is that the computational scheme is likely to be more stable as we are just computing many linear functions as opposed to carrying out many conditional maximizations.

Suppose now that we are primarily interested in the regression coefficient β_1. It is then preferable to work in a parameterization with β_1 the first component of θ, such as $\theta = (\beta_1, \beta_0, \phi)$. The marginal posterior density of β_1 may then be readily computed from (6). The density, represented by the dotted line in Figure 1, is hardly distinguishable from the exact (solid line).

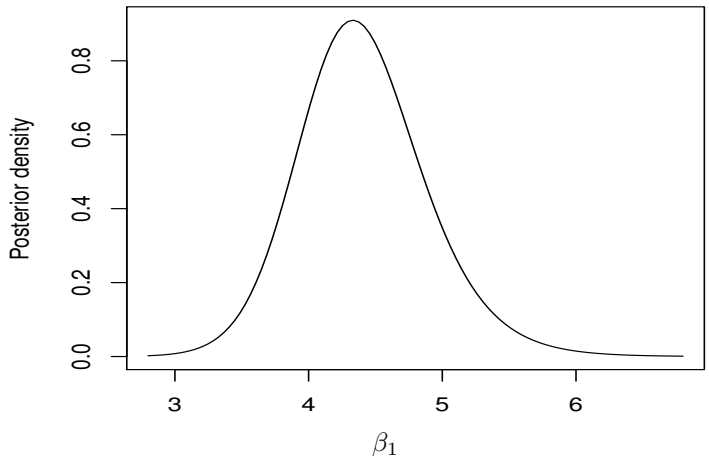

Figure 1: Marginal posterior density of β_1 for the censored normal regression model via importance sampling (dotted line) and exact (solid line).

We next apply the method of antithetic variates to this example. The signed root based simulation algorithm using exponential tilting was run with $m = 500$ and the posterior expectation of $v(\theta)$ was found to be -1.4982. We observe that the posterior expectation estimate shows a slight improvement, as well as a saving of cpu time, over those obtained using formula (5). The method of antithetic variates here provides a small but worthwhile reduction in variability.

4 Discussion

In this paper we have explored the use of importance sampling based on signed root log-likelihood ratios incorporating exponential tilting to obtain versions of various asymptotic formulae for

Bayesian computation. The twin advantages of using exponential tilting are computational stability and saving of computation time. In particular, we found that calculating signed roots using exponential tilting took about 1/10 of cpu time, and that generating a random sample of size $m = 1000$ via the signed root based algorithm achieves about a 60% time reduction. The computational stability aspect is particularly important in higher-dimensional models where repeated computation of conditional maxima may be difficult. In contrast, the tilting method just requires repeated computation of linear functions. The theory and examples indicate that there is good reason to believe that they will often perform extremely well in moderately-dimensioned problems. More extensive investigation is required, however, to study the full range of application and to make proper comparisons between various competing methods.

A set of [R] programs to implement the approach here is available on request. The main code is generic and the user only needs to supply the necessary code specific to his/her model.

References

1. Cerquetti, A. (2007). A note on Bayesian nonparametric priors derived from exponentially tilted Poisson-Kingman models. *Statistics and Probability Letters*, **77**, 1705–1711.
2. Crawford, D. E. (1970). Analysis of incomplete life test data on motorettes. *Insulation Circuits*, **16**, 43–48.
3. Evans, M. and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.
4. Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*. Methuen, London.
5. Kharroubi, S. A. and Sweeting, T. J. (2010). Posterior simulation via signed root log-likelihood ratios. *Bayesian Analysis*, **5**(4), 787–816.
6. Kharroubi, S. A. and Sweeting, T. J. (2016). Exponential tilting in Bayesian asymptotics. *Biometrika*, **103**(2), 337–349.
7. Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, **92**, 31-46.
8. Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood, *Annals of Statistics*, **35**, 634-672.
9. Schmee, J. and Hahn, G. J. (1979). A simple method for regression analysis with censored data. *Technometrics*, **21**, 417–432.
10. Sweeting, T. J. (1996). Approximate Bayesian computation based on signed roots of log-density ratios (with Discussion). *Bayesian Statistics, 5*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 427–444. Oxford University Press.
11. Sweeting, T. J. and Kharroubi, S. A. (2003). Some new formulae for posterior expectations and Bartlett corrections. *Test*, **12**, 497–521, 2003.
12. Sweeting, T. J. and Kharroubi, S. A. (2005). Application of a predictive distribution formula to Bayesian computation for incomplete data models. *Statistics and Computing*, **15**, 167–178.
13. Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.*, **82**, 528–540.