Comput Stat
DOI 10.1007/s00180-016-0677-z

ORIGINAL PAPER

# Integrated likelihood computation methods

**Zhenyu Zhao**[1] · **Thomas A. Severini**[1]

Received: 2 June 2015 / Accepted: 17 August 2016


**Abstract** Suppose a model has parameter $\theta = (\psi, \lambda)$, where $\psi$ is the parameter of interest and $\lambda$ is a nuisance parameter. The integrated likelihood method eliminates $\lambda$ from the likelihood function $L(\psi, \lambda)$ by integrating with respect to a weight function $\pi(\lambda|\psi)$. The resulting integrated likelihood function $\bar{L}(\psi)$ can be used for inference for $\psi$. However, the analytical form for the integrated likelihood is not always available. This paper discusses 12 different approaches to computing the integrated likelihood. Some methods were originally developed for other computation purposes and they are modified to fit in the integrated likelihood framework. Methods considered include direct numerical integration methods such as Monte Carlo integration method, importance sampling, Laplace method; marginal likelihood computation methods; and methods for computing the marginal posterior density. Simulation studies and real data example are presented to evaluate and compare these methods empirically.

**Keywords** MCMC · Likelihood inference · Bayesian · Nuisance parameter · Numerical integration

## 1 Introduction

Suppose a model has parameter $\theta \in \Theta$ and $\theta = (\psi, \lambda)$, where $\psi \in \Psi$ is the parameter of interest and $\lambda$ is a nuisance parameter. For a given $\psi \in \Psi$, let $\Lambda(\psi) = \{\lambda : (\psi, \lambda) \in \Theta\}$ denote the parameter space for $\lambda$ and the corresponding set of likelihoods can be expressed as $\mathcal{L}_\psi = \{L(\psi, \lambda) : \lambda \in \Lambda(\psi)\}$. Likelihood inference is often based on a pseudolikelihood function sumarizing $\mathcal{L}_\psi$. One well known approach is

✉ Zhenyu Zhao
zhenyuzhao2014@u.northwestern.edu

[1] Department of Statistics, Northwestern University, Evanston, IL, USA

Published online: 01 September 2016

the profile likelihood function, which summarizes $\mathcal{L}_\psi$ by its maximum value: $L_p(\psi) = L(\psi, \hat{\lambda}_\psi) = sup_{\lambda \in \Lambda(\psi)} L(\psi, \lambda)$. However, the profile likelihood method has several drawbacks. For instance, the profile likelihood may be misleading when the sample size is small or when the likelihood has a sharp "ridge"; see Berger et al. (1999) and Severini (2000, Ch. 4).

The integrated likelihood summarizes $\mathcal{L}_\psi$ by integration with a weight function $\pi(\lambda|\psi)$ for $\lambda$:

$$\bar{L}(\psi) = \int_\Lambda L(\psi, \lambda)\pi(\lambda|\psi)d\lambda. \tag{1}$$

See Berger et al. (1999), Kalbfleisch and Sprott (1970), Liseo (1993), Severini (2000, Ch. 8), and Severini (2007) for further discussion.

There are several possible advantages to the integrated likelihood. First, the integrated likelihood is based on integration rather than maximization and the average value is a often a better summary of $\mathcal{L}_\psi$ than is the maximum value. Second, the asymptotic properties of procedures based on an integrated likelihood are often superior to those based on the profile likelihood. Third, the integrated likelihood method is easily applied to certain nonregular models, in which other methods, such as those based on the profile, marginal, or conditional likelihood cannot be applied. See, e.g., Berger et al. (1999), Severini (1999), Severini (2007), Severini (2010) and Severini (2011) for further discussion.

Because any weight function $\pi(\lambda|\psi)$ can be used to define an integrated likelihood, selection of an appropriate weight function is an important issue. Berger et al. (1999) reviews different choices of weight functions and discusses their strengths and weaknesses. Liseo (1993) provides examples which suggest that the reference prior is preferable over the Jeffreys prior. Severini (2007) approaches this issue by first introducing a nuisance parameter $\phi$, called the zero-score-expectation parameter, and then choosing a weight function for $\phi$ which does not depend on $\psi$. The resulting integrated likelihood is closely related to the modified profile likelihood given by Barndorff-Nielsen (1980, 1983, 1995) and Severini (1998). The proposed integrated likelihood is also score-unbiased and information-unbiased to order $O(n^{-1})$, as well as parameterization-invariant.

An important difficulty in implementing the integrated likelihood method is computation complexity. The integration needed to obtain the integrated likelihood is generally not solvable by analytical methods. Even if it is solvable, it may be very complicated. Therefore, it is often necessary to find an efficient method to estimate the integral used to define the integrated likelihood. Though there are many methods available for estimating integrals, currently there is sparse literature developed specifically for the integrated likelihood.

The goal of this paper is to consider the application of different integration algorithms to the calculation of the integrated likelihood and to evaluate them using a simulation study. The methods covered include three categories: (1) direct approaches, including simple Monte Carlo, importance sampling, and Laplace's method; (2) methods developed originally for marginal likelihood (normalizing constant) in Bayesian inference, including the harmonic mean (Newton and Raftery 1994), Chib method

using Gibbs sampler (Chib 1995), Chib–Jeliazkov method using Metropolis–Hastings algorithm (Chib and Jeliazkov 2001), the posterior kernel density, and the bridge sampling method (Meng and Schilling 2002; Meng and Wong 1996; Chen 2005); and (3) methods developed originally for calculating a marginal posterior density, including the posterior kernel, the conditional marginal density estimator (CMDE) (Gelfand et al. 1992), and the importance weighted marginal density estimator (IWMDE) (Chen 1994). Since these methods are not developed for the integrated likelihood originally, some of them need modification to fit in this new context.

Bos (2002) evaluated marginal likelihood estimation methods under a simple regression setting. Bos (2002) concluded that both the Monte Carlo method and harmonic mean method are not trustworthy. The Laplace's method and kernel proved to be more useful. The Chib method can give good results, provided that computational costs are high especially with increasing dimension of the vector of parameters. Bos (2002) evaluated these methods in the setting of normal posterior, while in this paper, these methods are examined with more skewed distributions for calculating integrated likelihood.

This paper is arranged as follows. Section 2 discusses differences of the integration between integrated likelihood and Bayesian approach. Section 3 reviews different methods for computing the integrated likelihood. Section 4 presents the performance of the methods in simulation settings. Section 5 evaluates the methods with a real data example on NBA data.

## 2 Differences between computing an integrated likelihood and Bayesian computation

Although there are some similarities between integrated likelihood and Bayesian integration problems, they are different from each other in many ways.

In Bayesian approaches, the outcome of a marginal likelihood calculation is a normalizing constant and the outcome of a marginal posterior density calculation is a density function, while the integrated likelihood produces a pseudolikelihood. The integrated likelihood is different from the marginal likelihood, since the integrated likelihood is a function of $\psi$ and in general the integrated likelihood needs to be calculated at multiple $\psi$ values. Unlike the marginal posterior density, the integrated likelihood is not a density, thus it can be calculated up to a multiplying constant.

In particular, the marginal posterior density $p(\psi|y)$ must be integrated to one, while the integrated likelihood $\bar{L}(\psi)$ is allowed to have infinite integral. However, if $\bar{L}(\psi)$ has infinite integral, the methods originally developed for calculating marginal posterior density will not work. Instead, we could use the methods designed for calculating a marginal likelihood to calculate the integrated likelihood. This idea is consistent with Kass et al. (1998)'s statement, which indicated that when one model has an improper posterior distribution, then one solution is to perform inference conditional on the poorly identified parameters, rather than averaging over them. Hence, similarly, if the $\bar{L}(\psi)$ is not integrable, then $\bar{L}(\psi)$ can be calculated using methods which are conditional on $\psi$.

The weight function $\pi(\lambda|\psi)$ used in defining the integrated likelihood does not need to be a genuine density. The weight function is selected in order to yield an integrated

likelihood that is useful for non-Bayesian inference. We can choose $\pi(\lambda|\psi)$ to make computation easy, provided that the integrated likelihood has desired frequentist properties. Hence, the guidence for selecting a weight function is different from selecting a prior in Bayesian approach. Although the weight function is neither a genuine density nor a Bayesian prior, for convenience, we will refer to it as a "conditional prior".

## 3 Integration algorithms

This section describes the different computational algorithms that will be considered in this paper.

### 3.1 Direct approaches

#### *3.1.1 Simple Monte Carlo*

If the conditional prior $\pi(\lambda|\psi)$ is a genuine density function from which a sample $\{\lambda_i\}_{i=1}^n$ can be directly generated, then the simple Monte Carlo estimator is $\hat{\bar{L}}(\psi) = \frac{1}{n}\sum_{i=1}^n L(\psi, \lambda_i)$. However, if the prior is flat and corresponds to a uniform distribution, though the sample can be generated, the support of the uniform distribution needs to be carefully defined. If the range is too narrow, the estimation may be poor due to excluding possible $\lambda$ values. If the range is too wide, the estimation procedure may be inefficient since many points will be located in the low-likelihood region. Another issue to be considered is when the integrated likelihood $\bar{L}(\psi)$ needs to be estimated at different $\psi$, then the support range should be updated for each $\psi$, which is inconvenient.

On the other hand, if the conditional prior $\pi(\lambda|\psi)$ is not a genuine density function, then the sample can be generated by rejection sampling; see for example Givens and Hoeting (2005, Ch 6) for more details about rejection sampling.

The idea of Monte Carlo integration is simple and in many cases it is easy to implement. However, as noted by Bos (2002), the estimates can be very unstable and inefficient. Overall, the simple Monte Carlo integration algorithm can be summarized as

> Simple Monte Carlo algorithm:
>
> 1. Draw samples $\{\lambda_i\}_{i=1}^n$ from $\pi(\lambda|\psi)$, by direct sampling or rejection sampling.
> 2. Estimate the integrated likelihood by $\hat{\bar{L}}(\psi) = \frac{1}{n}\sum_{i=1}^n L(\psi, \lambda_i)$

#### *3.1.2 Importance sampling*

Importance sampling is similar to Monte Carlo integration except that the samples $\{\lambda_i\}_{i=1}^n$ are drawn from an *importance density* rather than from $\pi(\lambda|\psi)$. Thus, the first step of importance sampling is to choose an importance density $\pi^*(\lambda|\psi)$. If $\pi^*(\lambda|\psi)$

is chosen to approximate the conditional prior density $\pi(\lambda|\psi)$, then many draws will fall in a region in which the likelihood is relatively low. On the other hand, if $\pi^*(\lambda|\psi)$ is chosen to approximate the shape of the weighted likelihood function $L(\psi, \lambda)\pi(\lambda|\psi)$, then more draws will fall in a high-likelihood region, but it may result in an estimator with infinite variance (Newton and Raftery 1994).

Bos (2002) suggests using an importance density of the form $\delta\pi(\lambda|\psi) + (1 - \delta)L(\psi, \lambda)\pi(\lambda|\psi)$ ($\delta \in [0, 1]$) to yield a consistent estimate and to achieve better convergence performance; see also Newton and Raftery (1994). In the present context, ideally, the weight function $\pi(\lambda|\psi)$ does not have much influence on the resulting integrated likelihood function. Hence it is natural and simpler to define an importance density based only on the likelihood function $L(\psi, \lambda)$. For example, we can define the importance density as $\mathcal{N}(\hat{\lambda}_\psi, \hat{\Sigma}_{\lambda\lambda|\psi})$, where $\hat{\lambda}_\psi$ is the maximum likelihood estimate of $\lambda$ given $\psi$, $\hat{\Sigma}_{\lambda\lambda|\psi} = (-l_{\lambda\lambda}(\psi, \hat{\lambda}_\psi))^{-1}$ is the observed Fisher information, and $\mathcal{N}(\mu, \Sigma)$ denotes the density of the multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$.

The importance sampling procedure may be described as follows

Importance Sampling algorithm:

1. Draw $\{\lambda_i\}_{i=1}^n$ from an importance distribution with probability density function as $\pi^*(\lambda|\psi)$. One choice of importance distribution is $\mathcal{N}(\hat{\lambda}_\psi, \hat{\Sigma}_{\lambda\lambda|\psi})$, where $\hat{\lambda}_\psi = \text{argmax}_\lambda L(\psi, \lambda)$ and $\hat{\Sigma}_{\lambda\lambda|\psi}$ is the observed Fisher information.
2. The importance sampling estimator is

$$\hat{\hat{L}}(\psi) = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i L(\psi, \lambda_i),$$

where weight $w_i = \pi(\lambda_i|\psi)/\pi^*(\lambda_i|\psi)$.

### *3.1.3 Laplace's method*

The Laplace estimator for integrated likelihood is given by

$$\hat{\hat{L}}(\psi) = (\det \hat{\Sigma}_{\lambda\lambda|\psi})^{1/2} L(\psi, \hat{\lambda}_\psi)\pi(\hat{\lambda}_\psi|\psi).$$

where $\det \hat{\Sigma}_{\lambda\lambda|\psi}$ denotes the matrix determinant of observed Fisher information.

It is based on the idea that, for a large sample size, the values of $\lambda$ that make the most contribution to the integrated likelihood integral (1) are those near $\hat{\lambda}_\psi$; see Tierney and Kadane (1986), Barndorff-Nielsen and Nielsen (1989, Section 3.3), (Jensen 1995, Sect. 3.1), and Kass et al. (1990) for further details about Laplace approximations.

Unlike the Monte Carlo method and importance sampling, the Laplace approximation has an explicit form and does not use type of sampling. When the required maximization is easy to carry out, the Laplace approximation can be much faster

than methods based on sampling. The drawback of Laplace's method is that when the sample size is small, then the accuracy of the approximation is not guaranteed. Another drawback is that, since the Laplace approximation is based on maximization, in models when maximization is problematic, the Laplace-approximated integrated likelihood will have the same difficulties as the profile likelihood.

### 3.2 Methods from marginal likelihood estimation

In Bayesian analysis, the marginal likelihood, or equivalently the normalizing constant, is defined as

$$m(y) = \int L(\theta)\pi(\theta)d\theta,$$

where $L(\theta)$ denotes the likelihood and $\pi(\theta)$ is the prior density.

This is similar to the integrated likelihood situation in that they both integrate the likelihood function with a prior (or weight) function with respect to the parameter. Hence, the methods from marginal likelihood estimation can be adapted to estimate the integrated likelihood. The transformation from the marginal likelihood to the integrated likelihood problem is straightforward: for each given $\psi$, we treat $\psi$ as a fixed parameter and, given a choice of $\pi(\lambda|\psi)$, calculation of the integrated likelihood for $\psi$ is identical to a marginal likelihood calculation.

However, it is important to note that there are two differences between a marginal likelihood and an integrated likelihood. One is that the integration in the marginal likelihood calculation involves the entire parameter, while the integration in the integrated likelihood calculation involves only the nuisance parameter. Therfore, for a given data set, the marginal likelihood only has one value while the integrated likelihood must be evaluated at different $\psi$ values. Another difference is that there are fewer requirements on the conditional prior used in an integrated likelihood calculation than there are in a Bayesian analysis. Specifically, in the integrated likelihood context, the primary requirement is simply that the resulting integrated likelihood is useful for non-Bayesian inference and the conditional prior may be improper or the integral of the integrated likelihood with respect to $\psi$ may be infinite.

In this section, we consider modifications to different algorithms for estimating the marginal likelihood in order to apply them to calculation of an integrated likelihood.

#### *3.2.1 Harmonic mean*

Newton and Raftery (1994) developed the harmonic mean estimator for the marginal likelihood. Similarly, we can derive the harmonic mean estimator for integrated likelihood. Suppose that the conditional prior density for $\lambda$ is proper, so that

$$\int \pi(\lambda|\psi)d\lambda = 1 \quad \text{for all} \quad \psi.$$

According to the harmonic mean identity in Newton and Raftery (1994), the following harmonic mean identity holds for integrated likelihood:

$$\frac{1}{\bar{L}(\psi)} = \int \frac{p(\lambda|\psi, y)}{L(\psi, \lambda)} d\lambda = E\left(\frac{1}{L(\psi, \lambda)}|\psi, y\right). \tag{2}$$

Here $p(\lambda|\psi, y)$ denotes the conditional posterior density of $\lambda$ given $\psi$ based on $\pi(\lambda|\psi)$.

Then given samples $\{\lambda_i\}_{i=1}^n$ from the posterior density $p(\lambda|\psi, y)$, the harmonic mean estimator of $\bar{L}(\psi)$ can be defined as follows:

$$\hat{\bar{L}}_{HM}(\psi) = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{L(\psi, \lambda_i)}\right]^{-1}. \tag{3}$$

One advantage of the harmonic mean estimator is simplicity. Also, it is unbiased and consistent as the simulation size increases. However, the reciprocal of the likelihood can have infinite variance and hence is unstable (Raftery Adrian 2006). Raftery Adrian (2006) describe two ways to stablize the harmonic mean estimator. The first approach utilizes a technique of reducing the dimension of the parameter space, when some conditions are satisfied. The second approach is based on the fact that when the sample size underlying the likelihood increases to infinity, asymptotically, the posterior distribution of the loglikelihoods can be approximated by a shifted Gamma distribution. Given this asymptotic distribution, Raftery Adrian (2006) develop a way to approximate the harmonic mean using the moment generating function.

### *3.2.2 Chib method*

One well-known marginal likelihood estimator is Chib method using the Gibbs sampler (Tanner and Wong 1987 and Wei and Tanner 1990) output (Chib 1995), and this method is also applicable to computation of the integrated likelihood. The Chib's method of marginal likelihood computation is based on basic marginal likelihood identity (BMI). Similar to the Chib's BMI, the integrated likelihood identity can be defined as

$$\bar{L}(\psi) = \frac{L(\psi, \lambda)\pi(\lambda|\psi)}{p(\lambda|\psi, y)}, \tag{4}$$

that holds for any $\lambda \in \Lambda$. This identity can be derived in following steps:

$$\begin{aligned} p(\lambda|\psi, y) &= \frac{p(\psi, \lambda|y)}{p(\psi|y)} = \frac{p(\psi, \lambda|y)}{\int p(\psi, \lambda|y)d\lambda} = \frac{L(\psi, \lambda)\pi(\lambda|\psi)}{\int L(\psi, \lambda)\pi(\lambda|\psi)d\lambda} \\ &= \frac{L(\psi, \lambda)\pi(\lambda|\psi)}{\bar{L}(\psi)}. \end{aligned}$$

Therefore, the integrated likelihood identity reduces the integration calculation to density estimation of $p(\lambda|\psi, y)$. One related identity proposed by Chib (1995) is a

variant of BMI for random effects models, that has a latent parameter in the model in addition to the model parameters that wants to eliminate. The integrated likelihood identity is similar to that identity in terms of including a parameter $\psi$ in the calculation, but different in that the output for integrated likelihood is a function of $\psi$.

Taking logs of both sides of the integrated likelihood identity yields

$$\log \bar{L}(\psi) = \log L(\psi, \lambda) + \log \pi(\lambda|\psi) - \log p(\lambda|\psi, y). \tag{5}$$

To implement the Gibbs sampler, suppose $z$ is a latent variable and that the conditional density functions $p(\lambda|\psi, y, z)$ and $p(z|\lambda, \psi, y)$ are known and can be directly sampled from. Under these assumptions, samples $\{\lambda_i, z_i\}_{i=1}^n$ from the densities $p(\lambda|\psi, y)$ and $p(z|\psi, y)$ can be drawn using a Gibbs sampler.

Since

$$p(\lambda|\psi, y) = \int p(\lambda|\psi, y, z)p(z|\psi, y)dz,$$

the Monte Carlo estimator of $p(\lambda|\psi, y)$ based on the Gibbs output is

$$\hat{p}(\lambda|\psi, y) = \frac{1}{n}\sum_{i=1}^{n} p(\lambda|\psi, y, z_i).$$

By substituting this estimator in the integrated likelihood expression, the estimator of the integrated likelihood is given by:

$$\log \hat{\bar{L}}(\psi) = \log L(\psi, \lambda^*) + \log \pi(\lambda^*|\psi) - \log\left(\frac{1}{n}\sum_{i=1}^{n} p(\lambda^*|\psi, y, z_i)\right); \tag{6}$$

here $\lambda^*$ is a high density point which may depend on $\psi$, such as maximum likelihood estimate of $\lambda$ for fixed $\psi$.

As proposed by Chib (1995), when the nuisance parameter $\lambda$ is a vector, the Gibbs sampler can be applied to sample from conditional distributions of sub-blocks of $\lambda$, that makes the Chib's method work without latent variable.

The derived integrated likelihood estimator is simulation consistent for fixed $\psi$. In addition, this method uses the sample density average rather than the reciprocal of the likelihood; hence it is stable and does not suffer from the infinite variance problem present in the harmonic mean method. However, one difficulty of this procedure is that, in general integrated likelihood settings, the conditional distribution $p(\lambda|\psi, y, z)$ and $p(z|\lambda, \psi, y)$ are unknown or cannot be directly sampled from. Another issue is that the latent variable $z$ may be hard to develop. These drawbacks restrict the generality of the application of the Chib method to the integrated likelihood estimation.

To summarize, assume $p(\lambda|\psi, y, z)$ and $p(z|\lambda, \psi, y)$ are known. Then the Chib method procedure is as follows:

Chib method:

1. Run the Gibbs sampler. Set an initial value $\lambda_0$ (such as $\hat{\lambda}_\psi$) and set $i = 1$.
   - Draw $z_i \sim p(z|\lambda_{i-1}, \psi, y)$
   - Draw $\lambda_i \sim p(\lambda|\psi, y, z_i)$.
   - Record $\{\lambda_i, z_i\}$ in the sample set and update $i = i + 1$.
   - Update $i = i + 1$. Repeat these steps until the desired sample size is achieved. The generated samples are denoted by $\{\lambda_i, z_i\}_{i=1}^n$.
2. By the integrated likelihood identity (5), estimate the integrated likelihood by

$$\hat{\tilde{L}}(\psi) = \exp\left\{\log L(\psi, \lambda^*) + \log \pi(\lambda^*|\psi) - \log\left(\frac{1}{n}\sum_{i=1}^{n} p(\lambda^*|\psi, y, z_i)\right)\right\},$$

where $\lambda^*$ is a high density point, such as $\hat{\lambda}_\psi$.

### *3.2.3 Chib–Jeliazkov method*

Based on the basic marginal likelihood identity, Chib and Jeliazkov (2001) developed a less demanding and more general approach to estimating the posterior density and marginal likelihood, using Metropolis–Hastings sample output. This Chib–Jeliazkov method is a M-H variant of the Chib method, but more general than Chib method in the sense that it does not require creation of latent variable $z$ and the knowledge of the conditional density $p(\lambda|\psi, y, z)$ and $p(z|\lambda, \psi, y)$.

The Chib–Jeliazkov method is easily adapted to estimating integrated likelihood by performing the entire process conditional on a given $\psi$ value. Let $q(\lambda, \lambda'|\psi, y)$ denote the proposal density for the transition from $\lambda$ to $\lambda'$, conditional on the parameter value $\psi$ and let

$$\alpha(\lambda, \lambda'|\psi, y) = min\left\{1, \frac{L(\lambda', \psi)\pi(\lambda'|\psi)}{L(\lambda, \psi)\pi(\lambda|\psi)}\frac{q(\lambda', \lambda|\psi, y)}{q(\lambda, \lambda'|\psi, y)}\right\}$$

denote the probability of accepting the proposed value. Samples $\{\lambda_i\}_{i=1}^n$ can be drawn from the posterior density $p(\lambda|\psi, y)$ using Metropolis–Hastings sampling algorithm; see, for example, Chib and Greenberg (1995); Hastings (1970) and Tanner (1996). Samples $\{\lambda_j\}_{j=1}^J$ from $q(\lambda^*, \lambda|\psi, y)$ can be generated by direct sampling. Based on the results of Chib and Jeliazkov (2001), it is straightforward to show $p(\lambda^*|\psi, y)$ can be approximated by

$$\hat{p}(\lambda^*|\psi, y) = \frac{n^{-1}\sum_{i=1}^{n}\alpha(\lambda_i, \lambda^*|\psi, y)q(\lambda_i, \lambda^*|\psi, y)}{J^{-1}\sum_{j=1}^{J}\alpha(\lambda^*, \lambda_j|\psi, y)}. \tag{7}$$

Given this conditional estimator, the integrated likelihood estimator is

$$\hat{\tilde{L}}(\psi) = \exp(\log L(\psi, \lambda^*) + \log \pi(\lambda^*|\psi) - \log\left(\frac{n^{-1}\sum_{i=1}^{n} \alpha(\lambda_i, \lambda^*|\psi, y) q(\lambda_i, \lambda^*|\psi, y)}{J^{-1}\sum_{j=1}^{J} \alpha(\lambda^*, \lambda_j|\psi, y)}\right).$$

There are two extensions of the above procedure given by Chib and Jeliazkov (2001). The first one deals with models with latent variables. The second one enables the procedure to handle models with multiple parameter blocks, since for many high-dimemsional problems, it will be convenient and computationally necessary to sample parameters in several smaller blocks. The same extensions can be applied to the integrated likelihood estimation procedure in a straightforward way.

The advantage of this procedure is that it can be applied to all integrated likelihood models. The reason is that this procedure only requires the likelihood function and the conditional prior density, which are always available. Though proposal densities and block schemes must be selected, this method is robust to different selections (Chib and Jeliazkov 2001). Another practical advantage is that once the programming code has been built for solving one integrated likelihood problem, then, with slight modification, the code can be applied to other integrated likelihood problems.

In practice, if the simulations are repeated multiple times, then the empirical standard error can be calculated over those estimates. Chib and Jeliazkov (2001) also developed an approach to computing the numerical standard error of the estimator based on samples from one simulation.

To summarize, the estimation procedure can be implemented in the following steps:

Chib–Jeliazkov method:

1. Choose a proposal distribution with density $q(\lambda, \lambda'|\psi, y)$. The algorithm is generic that is not restricted for specific proposals. Example proposal distributions could be a tailored proposal, such as $\mathcal{N}(\hat{\lambda}_\psi, c \cdot \hat{\Sigma}_{\lambda\lambda|\psi})$; or a random walk proposal , such as $\mathcal{N}(\lambda, c \cdot \hat{\Sigma}_{\lambda\lambda|\psi})$, where $c$ is a constant (see Chib and Jeliazkov 2001).
2. Define the acceptance probability function

$$\alpha(\lambda, \lambda'|\psi, y) = \min\left\{1, \frac{L(\lambda', \psi)\pi(\lambda'|\psi)}{L(\lambda, \psi)\pi(\lambda|\psi)} \frac{q(\lambda', \lambda|\psi, y)}{q(\lambda, \lambda'|\psi, y)}\right\}$$

3. Run the Metropolis–Hastings algorithm. Set the initial value for $\lambda_0$, such as $\hat{\lambda}_\psi$. Set $i = 0$.
   - Generate one draw $\tilde{\lambda}$ from the proposal distribution with density $q(\lambda_i, \lambda|\psi, y)$.
   - Generate a random number $U$ from the uniform distribution [0, 1].
   - If $U < \alpha(\lambda_i, \tilde{\lambda}|\psi, y)$, take $\tilde{\lambda}$ into the sample $\lambda_{i+1} = \tilde{\lambda}$ ; otherwise, drop $\tilde{\lambda}$ and take $\lambda_{i+1} = \lambda_i$.
   - Update $i = i + 1$, and repeat this procedure until the target sample size $n$ is reached.

The resulting samples $\{\lambda_i\}_{i=1}^n$ are from the conditional posterior distribution $p(\lambda|\psi, y)$.

4. Draw samples $\{\lambda_j\}_{j=1}^J$ from the proposal distribution with density $q(\lambda^*, \lambda|\psi, y)$, where $\lambda^*$ could be any fixed value in the high density region, e.g. $\lambda^* = \hat{\lambda}_\psi$.
5. Using these samples, the estimate is given by

$$\hat{p}(\lambda^*|\psi, y) = \frac{n^{-1}\sum_{i=1}^n \alpha(\lambda_i, \lambda^*|\psi, y)q(\lambda_i, \lambda^*|\psi, y)}{J^{-1}\sum_{j=1}^J \alpha(\lambda^*, \lambda_j|\psi, y)}.$$

6. By the integrated likelihood identity (5), the estimate of the integrated likelihood is

$$\hat{\tilde{L}}(\psi) = \exp\{\log L(\psi, \lambda^*) + \log \pi(\lambda^*|\psi) - \log \hat{p}(\lambda^*|\psi, y)\}.$$

### *3.2.4 Posterior kernel estimator*

As discussed previously, the integrated likelihood identity reduces the problem of estimating the integrated likelihood into one of estimating the conditional posterior $p(\lambda^*|\psi, y)$; in addition, samples from this conditional posterior distribution can be drawn by a Gibbs sampler or by the Metropolis–Hastings procedure. Given the conditional posterior samples, a non-parametric way to estimate the density function is by kernel density estimator. There are several types of kernel functions that are commonly used (see Silverman 1986; Li and Racine 2007), and the posterior kernel algorithm is generic with respect to choice of kernel function. For following examples in this paper, the quartic (biweight) kernel function is used to illustrate the algorithm (though other kernels can also be applied), defined as

$$K(x) = \left(\frac{15}{16}\right)^d \Pi_{g=1}^d [1 - (x^{(g)})^2]^2 I(||x|| \leq 1),$$

where $x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$ is a $d$-dimensional variable and $||\cdot||$ denotes the Euclidean norm.

Using this posterior kernel method to estimate the integrated likelihood is relatively simple:

Posterior kernel estimator:

1. Generate draws $\{\lambda_i\}_{i=1}^n$ from the conditional posterior distribution with density $p(\lambda|\psi, y)$, by either a Gibbs sampler or the Metropolis–Hastings sampling algorithm.
2. Use a nonparametric kernel density estimator of $p(\lambda^*|\psi, y)$, given by

$$\hat{p}(\lambda^*|\psi, y) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\lambda^* - \lambda_i}{h}\right),$$

   where $h$ is the bandwidth that can be selected using cross-validation and $K()$ is the kernel function.
3. By the integrated likelihood identity (5), the integrated likelihood estimator is as follows:

$$\hat{\bar{L}}(\psi) = \exp\{\log L(\psi, \lambda^*) + \log \pi(\lambda^*|\psi) - \log \hat{p}(\lambda^*|\psi, y)\}.$$

### *3.2.5 Bridge sampling on marginal likelihood*

As discussed by Meng and Schilling (2002), bridge sampling (Meng and Wong 1996) can also be used to estimate the marginal likelihood, based on the identity proposed by Chib (1995).

Suppose $p_1(\theta)$ and $p_2(\theta)$ are two densities which are known up to a normalizing constant: $p_k(\theta) = q_k(\theta)/c_k,\ k = 1, 2$. Also assume that draws from these two densities are available. Bridge sampling provides a way to estimate the ratio of normalizing constants, given by $r = c_1/c_2$.

For the integrated likelihood, at any fixed $\psi$, the integrated likelihood $\bar{L}(\psi)$ can be viewed as the normalizing constant of the conditional posterior density $p(\lambda|\psi, y)$. That is, if we choose $p_1(\lambda|\psi) = p(\lambda|\psi, y)$ and $q_1(\lambda|\psi) = L(\psi, \lambda)\pi(\lambda|\psi)$, then

$$c_1(\psi) = \frac{q_1(\lambda|\psi)}{p_1(\lambda|\psi)} = \frac{L(\psi, \lambda)\pi(\lambda|\psi)}{p(\lambda|\psi, y)} = \bar{L}(\psi).$$

Following Meng and Schilling (2002), we can choose $p_2(\lambda|\psi) = q(\lambda^*, \lambda|\psi, y)$, which is the proposal density used in the Metropolis–Hastings algorithm. Since this assignment makes $p_2(\lambda|\psi)$ completely known, we can set $q_2(\lambda|\psi) = p_2(\lambda|\psi) = q(\lambda^*, \lambda|\psi, y)$, which leads to $c_2(\psi) = 1$. The ratio of interest becomes

$$r(\psi) = \frac{c_1}{c_2} = \bar{L}(\psi).$$

By Meng and Wong (1996), for any function $h(\psi, \lambda)$ which satisfies the non-zero finite integral condition

$$0 < |\int_{\Lambda} h(\psi,\lambda)p_1(\lambda|\psi)p_2(\lambda|\psi)d\lambda| < \infty,$$

we have the identity:

$$\bar{L}(\psi) = r(\psi) = \frac{c_1(\psi)}{c_2(\psi)} = \frac{\int_{\Lambda} h(\psi,\lambda)q_1(\lambda|\psi)p_2(\lambda|\psi)d\lambda}{\int_{\Lambda} h(\psi,\lambda)p_1(\lambda|\psi)q_2(\lambda|\psi)d\lambda} = \frac{E_2[q_1(\lambda|\psi)h(\psi,\lambda)]}{E_1[q_2(\lambda|\psi)h(\psi,\lambda)]},$$

where $E_k$ is taken with respect to $p_k$ for $k = 1, 2$. It follows that the bridge sampling estimator is

$$\hat{\bar{L}}(\psi) = \hat{r}(\psi) = \frac{\frac{1}{J}\sum_{j=1}^{J} q_1(\lambda_j|\psi)h(\psi,\lambda_j)}{\frac{1}{n}\sum_{i=1}^{n} q_2(\lambda_i|\psi)h(\psi,\lambda_i)},$$

where $\{\lambda_i\}_{i=1}^n$ are MCMC samples from $p(\lambda|\psi, y)$ and $\{\lambda_j\}_{j=1}^J$ are sampled directly from $q(\lambda^*, \lambda|\psi, y)$.

Note that for any function $h(\psi, \lambda)$ satisfying the non-zero finite integral condition, a bridge sampling estimator is defined. Under the assumption that all the samples are independent, Meng and Wong (1996) show that the optimal choice of $h(\psi, \lambda)$, which minimizes the asymptotic variance of $\log r(\psi)$ for fixed $\psi$, is

$$h^*(\psi,\lambda) \propto \frac{1}{s_1 q_1(\lambda|\psi) + s_2 r(\psi) q_2(\lambda|\psi)},$$

where $s_1 = n/(n + J)$ and $s_2 = J/(n + J)$. However, if the samples are not independent, as in our case, $h^*(\psi, \lambda)$ is no longer optimal. But Meng and Schilling (2002) indicate that the performance of the bridge sampling with $h^*(\psi, \lambda)$ is still good in many empirical situations unless the dependence among draws is too strong. For dealing with this dependence issue, Meng and Schilling (2002) suggest the use of the "effective size" $\tilde{n} = n(1 - \rho_1)/(1 + \rho_1)$ and $\tilde{J} = J(1 - \rho_2)/(1 + \rho_2)$ in defining $s_k$, where $\rho_k$ $(k = 1, 2)$ are estimated autocorrelation coefficients.

Another practical difficulty is that $h^*(\psi, \lambda)$ depends on the unknown function $r(\psi)$. For overcoming this issue, Meng and Wong (1996) developed the following iterative procedure

$$\hat{r}^*_{t+1}(\psi) = \frac{\frac{1}{J}\sum_{j=1}^{J} \frac{v(\psi,\lambda_j)}{s_1 v(\psi,\lambda_j) + s_2 \hat{r}^*_t(\psi)}}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{s_1 v(\psi,\lambda_i) + s_2 \hat{r}^*_t(\psi)}},$$

where $v(\psi, \lambda) = q_1(\lambda|\psi)/q_2(\lambda|\psi)$. For computational efficiency, it is suggested that the sequences $\{v(\psi, \lambda_i)\}_{i=1}^n$ and $\{v(\psi, \lambda_j)\}_{i=1}^J$ should be calculated before iterating. This iteration sequence, as shown by Meng and Wong (1996), has unique limit $\hat{r}^*(\psi)$ and $|\hat{r}^*_t(\psi) - \hat{r}^*(\psi)| \to 0$ monotonically in $t$. Also, the convergence rate is fast and 5 iterations generally work well in practice (Meng and Wong 1996). Hence, we can use this iterative procedure to obtain the optimal solution, without knowledge of the optimal $h^*(\psi, \lambda)$.

To summarize, the iterative bridge sampling estimation procedure is as follows:

Bridge sampling on marginal likelihood:

1. Generate $\{\lambda_i\}_{i=1}^n$ from $p(\lambda|\psi, y)$ using the Metropolis–Hastings sampling algorithm.
2. Generate $\{\lambda_j\}_{j=1}^J$ from the proposal density $q(\lambda^*, \lambda|\psi, y)$ which is defined in the Metropolis–Hastings sampling algorithm; $\lambda^*$ can be taken as $\hat{\lambda}_\psi$.
3. Determine the values of the sequences $\{v(\psi, \lambda_i)\}_{i=1}^n$ and $\{v(\psi, \lambda_j)\}_{j=1}^J$.
4. Run the iteration $n_c(\psi)$ times:

$$\hat{r}_{t+1}^*(\psi) = \frac{\frac{1}{J}\sum_{j=1}^J \frac{v(\psi,\lambda_j)}{s_1 v(\psi,\lambda_j)+s_2\hat{r}_t^*(\psi)}}{\frac{1}{n}\sum_{i=1}^n \frac{1}{s_1 v(\psi,\lambda_i)+s_2\hat{r}_t^*(\psi)}}.$$

Suppose that we are interested in the integrated likelihood evaluated at a sequence of $\psi$, denoted as $\{\psi_u\}_{u=1}^U$, and suppose $\psi_u$ and $\psi_{u+1}$ are close to each other for $u = 1, \ldots, U-1$. Then for estimating $\bar{L}(\psi)$ at $\psi = \psi_0$, we can take $\hat{r}_0^*(\psi_1) = 0$ as the initial value and $n_c(\psi_0) = 100$ as iteration times. For $\psi = \psi_u$ with $u > 1$, take $\hat{r}_0^*(\psi_u) = \hat{r}_{n_c(\psi_{u-1})}^*(\psi_{u-1})$ as the initial value and $n_c(\psi_u) = 5$ as the iteration times.

5. Take the estimate at the final iteration step as the estimated integrated likelihood:

$$\hat{\bar{L}}(\psi) = \hat{r}_{n_c(\psi)}^*(\psi).$$

### *3.2.6 Bridge sampling on posterior density*

An alternative bridge sampling approach to estimate the marginal likelihood is given by Chen (2005). Instead of applying the bridge sampling to the identity in Chib (1995); Chen (2005) notes that the bridge sampling can be used to estimate the posterior density $p(\theta|y)$, using the method given by Mira and Nicholls (2004). Since using the Chib (1995) identity reduces the problem of estimating the marginal likelihood to estimation of $p(\theta|y)$, this alternative bridge sampling method also leads to a marginal likelihood estimator.

For the integrated likelihood, the goal is to use bridge sampling to estimate the conditional posterior density $p(\lambda^*|\psi, y)$. Following Chen (2005) and Mira and Nicholls (2004), we define

$$\begin{aligned} p_1(\lambda|\psi) &= q(\lambda^*, \lambda|\psi, y), \\ q_1(\lambda|\psi) &= L(\psi, \lambda^*)\pi(\lambda^*|\psi)q(\lambda^*, \lambda|\psi, y), \\ c_1(\psi) &= q_1(\lambda|\psi)/p_1(\lambda|\psi) = L(\psi, \lambda^*)\pi(\lambda^*|\psi); \end{aligned}$$

and

$$\begin{aligned} p_2(\lambda|\psi) &= p(\lambda|\psi, y), \\ q_2(\lambda|\psi) &= L(\psi, \lambda)\pi(\lambda|\psi), \\ c_2(\psi) &= q_2(\lambda|\psi)/p_2(\lambda|\psi) = \bar{L}(\psi). \end{aligned}$$

Then the bridge sampling ratio is

$$r(\psi) = \frac{c_1(\psi)}{c_2(\psi)} = \frac{L(\psi, \lambda^*)\pi(\lambda^*|\psi)}{\bar{L}(\psi)} = p(\lambda^*|\psi, y).$$

Using these definitions, the regular bridge sampling steps can be implemented to obtain an estimator of $p(\lambda^*|\psi, y)$. The estimate of the integrated likelihood can be computed by substituting $\hat{p}(\lambda^*|\psi, y)$ in the integrated likelihood identity. This procedure can be summarized as follows:

Bridge sampling on posterior density:

1. Sample $\{\lambda_j\}_{j=1}^J$ from $q(\lambda^*, \lambda|\psi, y)$ directly.
2. Sample $\{\lambda_i\}_{i=1}^n$ from $p(\lambda|\psi, y)$ using the Metropolis–Hastings sampling algorithm.
3. Determine the value of the sequence $\{v(\psi, \lambda_j)\}_{j=1}^J$ and $\{v(\psi, \lambda_i)\}_{i=1}^n$, where function $v(\psi, \lambda)$ is defined as

$$v(\psi, \lambda) = q_1(\lambda|\psi)/q_2(\lambda|\psi) = \frac{L(\psi, \lambda^*)\pi(\lambda^*|\psi)q(\lambda^*, \lambda|\psi, y)}{L(\psi, \lambda)\pi(\lambda|\psi)}.$$

4. Run the following iteration

$$\hat{r}^*_{t+1}(\psi) = \frac{\frac{1}{n}\sum_{i=1}^n \frac{v(\psi,\lambda_i)}{s_1 v(\psi,\lambda_i)+s_2\hat{r}^*_t(\psi)}}{\frac{1}{J}\sum_{j=1}^J \frac{1}{s_1 v(\psi,\lambda_j)+s_2\hat{r}^*_t(\psi)}},$$

   where $s_1 = J/(n + J)$ and $s_2 = n/(n + J)$. The settings of initial values and number of iterations is the same as discussed in the previous section.
5. Take the estimate at the final iteration step as the estimated conditional posterior density

$$\hat{p}(\lambda^*|\psi, y) = \hat{r}^*_{n_c(\psi)}(\psi).$$

6. By the integrated likelihood identity (5), the estimator for integrated likelihood is given by

$$\hat{\bar{L}}(\psi) = \exp\{\log L(\psi, \lambda^*) + \log \pi(\lambda^*|\psi) - \log \hat{p}(\lambda^*|\psi, y)\}.$$

### 3.3 Methods based on estimation of the marginal posterior density

In a Bayesian framework, given a prior density $\pi(\theta)$, or equivalently $\pi(\psi, \lambda)$, let $p(\psi, \lambda|y)$ denote the posterior density of $(\psi, \lambda)$. Then the marginal posterior density of $\psi$ is defined as

$$p(\psi|y) = \int_{\Lambda(\psi)} p(\psi, \lambda|y) d\lambda.$$

Here $p(\psi, \lambda|y)$ is generally known up to an unknown normalizing constant $m(y)$. Since an MCMC sample can be obtained from $p(\psi, \lambda|y)$ without knowledge of $m(y)$, in general, this estimation problem can be solved by using an MCMC sample.

Note that the integrated likelihood estimation problem is quite similar to this marginal posterior density estimation problem in that both integrate over a subset of the entire parameter space of $\theta$. In fact, the marginal posterior density of $\psi$ can be written as

$$p(\psi|y) = \int_{\Lambda(\psi)} \frac{L(\psi, \lambda)\pi(\psi, \lambda)}{m(y)} d\lambda = \frac{\pi(\psi)}{m(y)} \int_{\Lambda(\psi)} L(\psi, \lambda)\pi(\lambda|\psi) d\lambda.$$

If $\pi(\lambda|\psi)$ is a proper weight function, then the relation between the marginal posterior density and the integrated likelihood function can be written as

$$p(\psi|y) = \frac{\pi(\psi)}{m(y)} \bar{L}(\psi).$$

However, there are some differences between marginal posterior density estimation and integrated likelihood estimation. First, in the Bayesian context a joint prior $\pi(\psi, \lambda)$ is assigned, but in the integrated likelihood approach, only a weight function $\pi(\lambda|\psi)$ for the nuisance parameter $\lambda$ is needed. Second, the goal of marginal posterior density estimation is estimation of a density, while the goal of integrated likelihood estimation is a pseudolikelihood function. That is, unlike marginal posterior density estimation, in the integrated likelihood case, the constant term $m(y)$ can be neglected and the resulting function does not require normalization. Hence the marginal posterior density and the integrated likelihood are different. However, note that, if the normalizing constant is neglected, if a flat prior is proposed for $\psi$, and, if a proper density function is used for $\pi(\lambda|\psi)$, then estimation of an integrated likelihood function is identical to estimation of a marginal posterior density.

Methods for estimating a marginal posterior density estimation are based on the samples from the posterior distribution $p(\psi, \lambda|y)$. Similarly, for estimating an integrated likelihood, samples $\{\psi_i, \lambda_i\}_{i=1}^n$ from the normalized weighted likelihood $L(\psi, \lambda|y)\pi(\lambda|\psi)/m(y)$ can be drawn by the MCMC sampling method, and these samples can be used for the estimation. The sampling scheme is different from the one used in estimation of a marginal likelihood. First, here both $\psi$ and $\lambda$ are sampled, while only samples of $\lambda$ are used in the previous section. In addition, in this section, only one sample set $\{\psi_i, \lambda_i\}_{i=1}^n$ is drawn and it is used for estimation of the integrated

likelihood evaluated at different $\psi$ values, while in the last section, one sample set $\{\lambda_i\}_i^n$ is drawn for each $\psi$ at which the integrated likelihood is evaluated.

The Metropolis–Hastings procedure can be used for sampling $\{\psi_i, \lambda_i\}_{i=1}^n$ from the normalized weighted likelihood $\frac{L(\psi,\lambda|y)\pi(\lambda|\psi)}{m(y)}$, as described below:

Metropolis–Hastings sampling algorithm:

1. Choose a proposal density $q(\psi, \lambda, \psi', \lambda'|y)$, which denotes the distribution density of $(\psi', \lambda')$ given the value $(\psi, \lambda)$ as the parameter value of the previous step. For example, $N[(\psi, \lambda), c \cdot \hat{\Sigma}_{\theta\theta})$ or $N[(\hat{\psi}, \hat{\lambda}), c \cdot \hat{\Sigma}_{\theta\theta})$ can be used as the proposal distribution in practice.
2. The acceptance probability function is defined as

$$\alpha(\psi, \lambda, \psi', \lambda'|y) = \min\{1, \frac{L(\psi', \lambda')\pi(\lambda'|\psi')}{L(\psi, \lambda)\pi(\lambda|\psi)} \frac{q(\psi', \lambda', \psi, \lambda|y)}{q(\psi, \lambda, \psi', \lambda'|y)}\}.$$

3. Run the Metropolis–Hastings algorithm with initial value $(\hat{\psi}, \hat{\lambda})$. The resulting sample is denoted as $\{\psi_i, \lambda_i\}_{i=1}^n$.

### *3.3.1 Posterior kernel estimator*

If the goal is to estimate the marginal posterior density $p(\psi|y)$, then $\{\psi_i\}_{i=1}^n$, drawn from this distribution, can be used to estimate this density directly using a kernel estimator. The same idea applies to the integrated likelihood. The kernel estimator for the integrated likelihood has the form

$$\hat{\tilde{L}}_{Kern}(\psi) = \frac{1}{nh^m} \sum_{i=1}^{n} K\left(\frac{\psi - \psi_i}{h}\right),$$

where $m$ is the dimemsion of $\psi$.

The advantage of the kernel estimator is that it does not require knowledge of the form of the target distribution, thus it is known as a "black-box" density estimator. However, the drawbacks are that the kernel method is generally inefficient for estimating a density value at a given point and the consistency of the estimator is unknown when the sample is drawn by the MCMC procedure. See Chen (2005) for more discussion.

### *3.3.2 Conditional marginal density estimator (CMDE)*

Assuming the density function $p(\psi|\lambda, y)$ is known, Gelfand et al. (1992) developed the CMDE method to calculate the marginal posterior density. This approach is easily modified to apply to estimation of the integrated likelihood function:

$$\begin{aligned}\bar{L}(\psi^*) &= \int_{\Lambda(\psi^*)} L(\psi^*, \lambda)\pi(\lambda|\psi^*)d\lambda \\ &= \int_{\Lambda(\psi^*)} \frac{L(\psi^*, \lambda)\pi(\lambda|\psi^*)}{\int_\Psi L(\psi, \lambda)\pi(\lambda|\psi)d\psi} \int_\Psi L(\psi, \lambda)\pi(\lambda|\psi)d\psi d\lambda \\ &= m(y)\int_\Psi \int_{\Lambda(\psi^*)} \frac{L(\psi^*, \lambda)\pi(\lambda|\psi^*)}{\int_\Psi L(\psi, \lambda)\pi(\lambda|\psi)d\psi} \frac{L(\psi, \lambda)\pi(\lambda|\psi)}{m(y)} d\psi d\lambda \\ &= m(y)\int_\Theta \frac{L(\psi^*, \lambda)\pi(\lambda|\psi^*)}{\int_\Psi L(\psi, \lambda)\pi(\lambda|\psi)d\psi} \cdot I(\lambda \in \Lambda(\psi^*)) \cdot \frac{L(\psi, \lambda)\pi(\lambda|\psi)}{m(y)} d\psi d\lambda.\end{aligned}$$

Under the assumption that $\Lambda(\psi)$ does not depend on $\psi$, the expression can be simplified as:

$$\bar{L}(\psi^*) = m(y)\int_\Theta \frac{L(\psi^*, \lambda)\pi(\lambda|\psi^*)}{\int_\Psi L(\psi, \lambda)\pi(\lambda|\psi)d\psi} \frac{L(\psi, \lambda)\pi(\lambda|\psi)}{m(y)} d\psi d\lambda.$$

The estimator of the integrated likelihood derived from this identity is given by

$$\hat{\bar{L}}(\psi^*) = \sum_{i=1}^{n} \frac{L(\psi^*, \lambda_i)\pi(\lambda_i|\psi^*)}{\int_\Psi L(\psi, \lambda_i)\pi(\lambda_i|\psi)d\psi}.$$

As shown by Gelfand et al. (1992), the advantage of this estimator is that it is unbiased and consistent. In addition, it is more efficient than the kernel estimator under a wide range of loss functions, in that it takes advantage of the known structure in the model.

However, the major drawback of this approach is that a closed form expression for $\int_\Psi L(\psi, \lambda)\pi(\lambda|\psi)d\psi$ is generally unavailable. The same difficulty arises for the CMDE, where the conditional posterior $p(\psi|\lambda, y)$ is unknown in most cases. The following approach by Chen (1994) is an extension to the CMDE designed to overcome this difficulty.

### *3.3.3 Importance weighted marginal density estimator (IWMDE)*

The IWMDE method proposed by Chen (1994) is a generalization of the CMDE that can be applied when $p(\psi|\lambda, y)$ not available. Given a known conditional density $w(\psi|\lambda)$ as the IWMDE weight function, we can write the integrated likelihood with parameter value $\psi^*$ as:

$$\begin{aligned}\bar{L}(\psi^*) &= \int_\Lambda L(\psi^*, \lambda)\pi(\lambda|\psi^*)d\lambda \\ &= \int_\Theta w(\psi|\lambda)L(\psi^*, \lambda)\pi(\lambda|\psi^*)d\psi d\lambda \\ &= m(y)\int_\Theta w(\psi|\lambda)\frac{L(\psi^*, \lambda)\pi(\lambda|\psi^*)}{L(\psi, \lambda)\pi(\lambda|\psi)} \frac{L(\psi, \lambda)\pi(\lambda|\psi)}{m(y)} d\psi d\lambda.\end{aligned}$$

Given a sample $\{\psi_i, \lambda_i\}_{i=1}^n$ from the normalized weighted likelihood $L(\psi, \lambda|y)\pi(\lambda|\psi)/m(y)$, the importance weighted integrated likelihood estimator (IWILE) is given by

$$\hat{\bar{L}}(\psi^*) = \frac{1}{n}\sum_{i=1}^{n} w(\psi_i|\lambda_i)\frac{L(\psi^*, \lambda_i)\pi(\lambda_i|\psi^*)}{L(\psi_i, \lambda_i)\pi(\lambda_i|\psi_i)}.$$

Following the guidelines given by Chen (1994), the weight function $w(\psi|\lambda)$ may be chosen as follows. Define the joint weight distribution as $\mathcal{N}((\bar{\psi}, \bar{\lambda}), \bar{\Sigma})$, where $(\bar{\psi}, \bar{\lambda})$ is the sample mean and $\bar{\Sigma}$ is the sample variance of $\{\psi_i, \lambda_i\}_{i=1}^n$. Note that $\theta = (\psi', \lambda')'$ so that $\bar{\Sigma}$ is given by

$$\bar{\Sigma} = \begin{pmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{21} & \bar{\Sigma}_{22} \end{pmatrix} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(\theta_i - \bar{\theta})'(\theta_j - \bar{\theta}).$$

The corresponding conditional distribution of $\psi$ given $\lambda$ is $\mathcal{N}(\tilde{\psi}, \tilde{\Sigma})$, where

$$\tilde{\psi} = \bar{\psi} + \bar{\Sigma}_{12}\bar{\Sigma}_{22}^{-1}(\lambda - \bar{\lambda}),$$

and

$$\tilde{\Sigma} = \bar{\Sigma}_{11} - \bar{\Sigma}_{12}\bar{\Sigma}_{22}^{-1}\bar{\Sigma}_{21}.$$

The weight function $w(\psi|\lambda)$ may then be defined as the density of $\mathcal{N}(\tilde{\psi}, \tilde{\Sigma})$.

Chen (2005) summarized several properties of the IWMDE estimator. First, this estimator is unbiased. Second, it is consistent under the ergodicity condition. Third, this estimator is better than the kernel method in terms of the Kullback–Leibler divergence.

## 4 Simulation study

In this section, we present three examples to evaluate the algorithms for computing integrated likelihood discussed in the previous section. All the computation was performed on a computation cluster, and each node on the cluster has a 260 AMD Opteron CPU core and 4 GB of memory. The time measured in the examples is CPU time measured on single node.

### 4.1 Example 1: Ratio of normal means

The ratio of normal means is a classical example in statistics. It is used in bioassay, among other places (Cox 1985; Srivastava 1986; Mendoza and ITAM, R. H., and Angel, S. 2005). It is also an example in which standard likelihood methods don't work very well and integrated likelihood methods tend to work better (Liseo 1993; Severini 2007).

In this example, we're looking at a simple case in which the standard deviations are taken to be 1 and $X$ and $Y$ are the sample means (so they have variance $1/N$ with $N$ denotes the sample size). Let $X$ and $Y$ denote two independent random variables, such that $X \sim \mathcal{N}(\mu_1, 1/N)$, $Y \sim \mathcal{N}(\mu_2, 1/N)$. Let $\psi = \mu_1/\mu_2$ be the parameter of interest and take $\lambda = \mu_2$ as the nuisance parameter. To specify the weight function, we use the parameterization $(\psi, \phi)$, where $\phi$ is called the "zero-score-expectation" parameter introduced by Severini (2007). By using this parameterization and putting a weight function $\pi(\phi)$ which is independent of $\psi$, the derived integrated likelihood has desirable frequency properties; see Severini (2007) for further details. For this example, the zero-score-expectation parameter is given by $\phi = (\psi^2+1)\lambda/(\psi\hat{\psi}+1)$. If the prior $\pi(\phi)$ is taken as $\mathcal{N}(0, 1/\tau^2)$, then the corresponding conditional prior for $\lambda$ given $\psi$ is $\mathcal{N}(0, (\psi\hat{\psi}+1)^2/[(\psi^2+1)\tau]^2)$.

For this choice, the integrated likelihood takes the form

$$\begin{aligned}\bar{L}(\psi;\tau^2) &= \int_{-\infty}^{\infty} L(\psi,\phi)\pi(\phi;\tau^2)d\phi \\ &= \int_{-\infty}^{\infty} \exp\left\{-\frac{N}{2}\frac{(\psi\hat{\psi}+1)^2}{\psi^2+1}(\phi^2-2y\phi)\right\}\exp\left\{-\frac{\tau^2\phi^2}{2}\right\}d\phi.\end{aligned}$$

The analytical closed form of the integrated log-likelihood is

$$\begin{aligned}\bar{l}(\psi;\tau^2) = \ & \frac{N}{2}\frac{(\psi x+y)^2}{\psi^2+1}\left[1+\frac{\tau^2}{N}\frac{\psi^2+1}{(\psi\hat{\psi}+1)^2}\right]^{-1} \\ & +\frac{1}{2}log(\psi^2+1) - log|\psi\hat{\psi}+1| + log\left[1+\frac{\tau^2}{N}\frac{\psi^2+1}{(\psi\hat{\psi}+1)^2}\right];\end{aligned}$$

see Severini (2007).

The maximum likelihood estimators of $\psi$ and $\phi$ are

$$\begin{aligned}\hat{\psi} &= \frac{x}{y} \\ \hat{\phi} = \hat{\lambda} &= \frac{x\hat{\psi}+y}{\hat{\psi}^2+1}.\end{aligned}$$

For the simulation study, we take $\tau = 1$ and consider the following integrated likelihood estimators.

**Class 1: Direct approaches introduced in Sect.** 3.1.

- $\hat{\bar{L}}_{MC}$: Simple Monte Carlo method.
- $\hat{\bar{L}}_{IS}$: Importance sampling method.
- $\hat{\bar{L}}_{Laplace}$: Laplace's method.

**Class 2: Methods introduced in Sect.** 3.2. These estimators use MCMC samples $\{\lambda_i\}_{i=1}^n$ from the conditional posterior $p(\lambda|\psi, y)$.

- $\hat{\hat{L}}_{HM}$: Harmonic mean estimator. Samples are generated by the Metropolis–Hastings algorithm.
- $\hat{\hat{L}}_{Chib}$: Chib method. For implementing Gibbs sampling, we need to introduce latent data, denoted as $z$. Let $z \sim \mathcal{N}(\lambda, 1/N)$. Denote the likelihood including latent data $z$ as $L^*(\psi, \lambda)$. Then the conditional posterior distribution of $\lambda$ is

$$\begin{aligned} p(\lambda|\psi, x, y, z) &\propto L^*(\psi, \lambda)\pi(\lambda|\psi) \\ &\propto \exp\left\{-\frac{N}{2}(z-\lambda)^2\right\} L(\psi, \lambda)\pi(\lambda|\psi) \\ &\propto \exp\left\{-\frac{\lambda^2}{2}\left[N(\psi^2+2) + \frac{(\psi^2+1)^2}{(\psi\hat{\psi}+1)^2}\right] + \lambda N(\psi x + y + z)\right\}. \end{aligned}$$

  That is, the conditional posterior distribution of $\lambda$ is normal distribution with mean $\frac{N(\psi x+y+z)}{N(\psi^2+2)+(\psi^2+1)^2/(\psi\hat{\psi}+1)^2}$ and variance $\frac{1}{N(\psi^2+2)+(\psi^2+1)^2/(\psi\hat{\psi}+1)^2}$.
- $\hat{\hat{L}}_{C-J}$: The Chib–Jeliazkov method using Metropolis–Hastings output with profile proposal distribution : $q(\lambda, \lambda'|\psi, y)$ that is the density function of $\mathcal{N}(\hat{\lambda}_\psi, \hat{\Sigma}_{\lambda\lambda|\psi})$ evaluating at point $\lambda'$.
- $\hat{\hat{L}}_{Kern-1}$: The kernel estimator using posterior sample $\{\lambda_i\}_{i=1}^n$ from the Metropolis–Hastings process with profile proposal distribution.
- $\hat{\hat{L}}_{BS-M}$: Bridge sampling method which is applied to the marginal likelihood equation.
- $\hat{\hat{L}}_{BS-P}$: Bridge sampling method which is applied to the conditional posterior equation.

**Class 3 with parameterization** $(\psi, \lambda)$: **Methods introduced in Sect.** 3.3. These estimators use samples $\{\psi_i, \lambda_i\}_{i=1}^n$ from the joint posterior $p(\psi, \lambda|y)$. The samples are generated by the Metropolis–Hastings procedure with proposal distribution defined as

$$\mathcal{N}(\begin{pmatrix}\hat{\psi}\\ \hat{\lambda}\end{pmatrix}, \hat{\Sigma}(\hat{\psi}, \hat{\lambda})),$$

where

$$\hat{\Sigma}(\psi, \lambda) = \begin{pmatrix} -l_{\psi\psi} & -l_{\psi\lambda} \\ -l_{\psi\lambda} & -l_{\lambda\lambda}\end{pmatrix}^{-1}.$$

- $\hat{\hat{L}}_{Kern-2L}$: The kernel estimator using only $\{\psi_i\}_{i=1}^n$ sample.
- $\hat{\hat{L}}_{IWMDE-L}$: The IWMDE estimator using $\{\psi_i, \phi_i\}_{i=1}^n$.

**Class 3 with zero score-expectation parameterization** $(\psi, \phi)$: **Methods introduced in Sect.** 3.3. These estimators use samples $\{\psi_i, \phi_i\}_{i=1}^n$ from the joint posterior $p(\psi, \phi|y)$. The samples are generated by the Metropolis–Hastings procedure with proposal distribution defined as

**Table 1** Empirical comparison of different computation methods in example 1

| Class | Method | Mean RMSE | CPU time (seconds per trial) |
|---|---|---|---|
| Class 1 | Monte Carlo (MC) | 1.720E−1 | 5.20 |
| | Importance sampling (IS) | 1.462E−1 | 15.07 |
| | Laplace's method (laplace) | 1.578E−6 | 2E−3 |
| Class 2 | Harmonic mean (HM) | 4.425E−1 | 51.24 |
| | Chib method (Chib) | 1.229E−4 | 25.42 |
| | Chib–Jeliazkov method (C–J) | 1.643E−5 | 110.28 |
| | Kernel with λ sample (Kern-1) | 1.282E−2 | 54.78 |
| | Bridge-marginal likelihood (BS-M) | 1.375E−5 | 147.75 |
| | Bridge-conditional posterior (BS-P) | 1.375E−5 | 163.24 |
| Class 3 $(\psi, \lambda)$ | Kernel with $\psi$ sample (Kern-2L) | 1.949E−1 | 21.93 |
| | IWMDE (IWMDE-L) | $1.985E1$ | 74.94 |
| Class 3 $(\psi, \phi)$ | Kernel with $\psi$ sample (Kern-2P) | 8.908E−1 | 20.39 |
| | IWMDE (IWMDE-P) | 1.073E−3 | 68.05 |

$$\mathcal{N}\left(\begin{pmatrix}\hat{\psi}\\ \hat{\phi}\end{pmatrix}, \hat{\Sigma}(\hat{\psi}, \hat{\phi})\right),$$

where

$$\hat{\Sigma}(\psi, \phi) = \begin{pmatrix} -l_{\psi\psi} & -l_{\psi\phi} \\ -l_{\psi\phi} & -l_{\phi\phi} \end{pmatrix}^{-1}.$$

- $\hat{\bar{L}}_{Kern-2P}$: The kernel estimator using only $\{\psi_i\}_{i=1}^n$ sample.
- $\hat{\bar{L}}_{IWMDE-P}$: The IWMDE estimator using $\{\psi_i, \phi_i\}_{i=1}^n$.

Suppose there are $K$ different $\psi$ points evaluated for each trial; then the simulation MSE of the estimator is defined as

$$MSE = \frac{1}{K}\sum_{k=1}^{K}\left[\log\left(\frac{\hat{\bar{L}}(\psi_k)}{\max_j(\hat{\bar{L}}(\psi_j))}\right) - \log\left(\frac{\bar{L}(\psi_k)}{\max_j(\bar{L}(\psi_j))}\right)\right]^2.$$

To perform the simulation study, we set $N = 10$, $\mu_1 = 4$ and $\mu_2 = 1/5$. Then a pair of $(x, y)$ is generated as $(4.1702, 0.1408)$. Based on these data, the integrated likelihood can be determined. As we want to calculate the integrated likelihood at different values of $\psi$, a sequence of $50\psi$ points is defined in the interval $[3.5, \hat{\psi} + 15]$. For each trial, the integrated likelihood at these $50\psi$ points will be calculated by different methods. The same procedure will be replicated 100 times. The empirical root-mean-squared-error (RMSE) of each estimator is calculated at each replicate, and the mean of the RMSE values is calculated using the 100 MSE values from these replications. For calculating each single value of the integrated likelihood, 10,000 simulation samples are used. The simulation results are shown in Table 1 and Fig. 1.

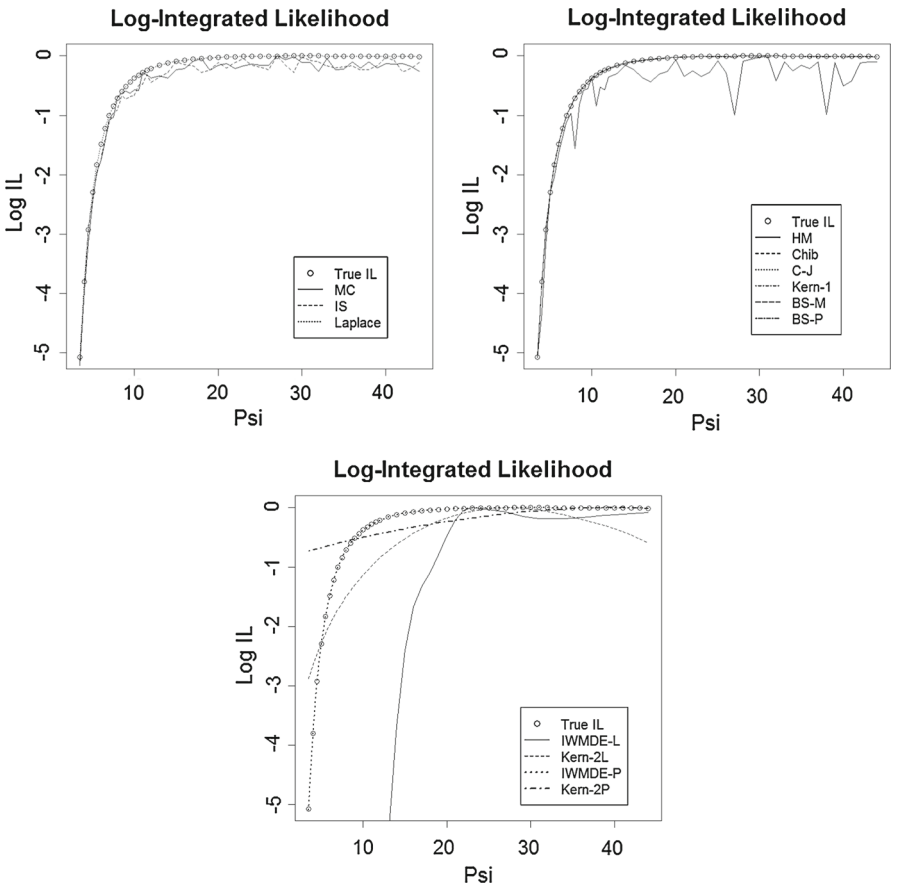


**Fig. 1** Example 1 Class 1, Class 2, and Class 3 estimators

The Laplace estimator has the smallest average RMSE among all algorithms. The reason is that in this example, the weighted likelihood of $\lambda$ takes the form of normal density. All Chib method, Chib–Jeliazkov method, and two Bridge Sampling methods can be regarded as the second best. In this example, the simulation difference between the two Bridge Sampling methods is negligible. It is also worth noting that the parameterization $(\psi, \phi)$ can hugely improve the performance of IWMDE method. The reason is that the IWMDE method uses the same sample set for calculating $\bar{L}(\psi)$ at different $\psi$ points. While the sample set is good for calculating $\bar{L}(\psi)$ at a point around $\hat{\psi}$, for a point far away from $\hat{\psi}$, then $\hat{\lambda}_\psi$ can be quite different from $\hat{\lambda}$, and the sample set will be poor for calculation. But the zero-score-expectation parameter $\phi$ has the property that $\hat{\phi}_\psi = \hat{\phi} + O(n^{-1/2})O(|\psi - \hat{\psi}|)$; hence even for $\psi$ far from $\hat{\psi}$, the sample set will still be yield accurate estimates. For execution time, the Laplace also takes the shortest time, since it does not need any samples. The other two direct approaches also run fast, however, they are not accurate. For the Class 2 estimators, the Chib method is the fastest estimator. The harmonic mean algorithm is also fast,

since it has a simple formula form. All other estimators within Class 2 take a relatively long time to run. Estimators in Class 3 have the feature that they only sample once for each trial, which can save some time. But for $\bar{l}_{IWMDE}$, a relatively large proportion of time will be spent on calculating the weight function.

### 4.2 Example 2: Normal distributions with common mean

Consider independent variables $Y_{jk} \sim \mathcal{N}(\psi, \lambda_j)$, where $\lambda_j > 0, k = 1, \ldots, N_j, j = 1, \ldots, m$. The zero score-expectation parameter is $\phi = (\phi_1, \ldots, \phi_m)$, where $\phi_j = \lambda_j - (\hat{\psi} - \psi)^2$ for $j = 1, \ldots, m$ Severini (2007, Example 4). If the weight function is defined as $\pi(\phi) = 1$ with support $(0, +\infty)$, then equivalently, the weight function for $\lambda$ will be $\pi(\lambda|\psi) = 1$ with support $((\hat{\psi} - \psi)^2, +\infty)$. The integrated likelihood will be

$$\bar{L}(\psi) = \Pi_{j=1}^{m} \int_{(\hat{\psi}-\psi)^2}^{\infty} \Pi_{k=1}^{N_j} \frac{1}{\sqrt{2\pi\lambda_j}} exp\left\{-\frac{(y_{jk} - \psi)^2}{2\lambda_j}\right\} d\lambda_j.$$

The analytical closed-form of the integrated likelihood function is given by

$$\bar{L}(\psi) = \Pi_{j=1}^{m} \frac{1}{s_j(\psi)^2}[1 - exp\{s_j(\psi)^2/(\hat{\psi} - \psi)^2\}], \tag{10}$$

where

$$s_j(\psi) = \sum_{k=1}^{n_j} (y_{jk} - \psi)^2.$$

All the algorithms except Chib method mentioned in previous section can be directly applied to this example. Though we can use the sub-block sampling approach for Chib method, using the fact that the conditional distribution of each $\lambda$ dimention is Inverse Gamma. However, this knowledge is equivalent to knowing the analytical form of integrated likelihood. So we evaluate the performance of other methods without this key information. For the Monte Carlo estimator, we use the uniform distribution as the random number generator. The range of the uniform distribution for generating the sample of $\lambda_j$ is $[(\hat{\psi} - \psi)^2, (\hat{\psi} - \psi)^2 + \hat{\lambda}_{\psi j} + 10 \times (-l_{\lambda_j \lambda_j}(\psi, \hat{\lambda}_{\psi}))^{-1/2}]$. For the importance sampling estimator, we still use the normal proposal, but if a $\lambda$ sample point with negative component is generated, then the weight for the sample point is taken to 0. For all other estimators of Class 2, simulation samples are generated by the Metropolis–Hastings algorithm. The proposal distribution used in the Metropolis–Hastings algorithm is the same uniform distribution as the simple Monte Carlo case. For Class 3 estimators, the Metropolis–Hastings method within the Gibbs algorithm is used to generate the simulation sample. The sample of $\psi$ is generated from a normal distribution with mean $\hat{\psi}$ and variance $(\sum_{j=1}^{m} N_j \lambda_{jt})^{-1}$, where $\lambda_t$ is the sample value of $\lambda$ at current step. Given the current simulation sample value of $\psi$ the sample of $\lambda$ is generated using the Metropolis–Hastings algorithm similar to what is done in the

**Table 2** Empirical comparison of different computation methods in example 2

| Class | Method | Mean RMSE | CPU time (seconds per trial) |
|---|---|---|---|
| Class 1 | Monte Carlo (MC) | 2.956 | 123.00 |
| | Importance sampling (IS) | 3.834 | 213.57 |
| | Laplace's method (laplace) | 4.198E−1 | 0.03 |
| Class 2 | Harmonic mean (HM) | 3.379 | 581.02 |
| | Chib–Jeliazkov method (C–J) | 7.429E−2 | 1281.59 |
| | Kernel with $\lambda$ sample (Kern-1) | 2.354 | 473.69 |
| | Bridge-Marginal likelihood (BS-M) | 1.130E−1 | 1163.56 |
| | Bridge-Conditional posterior (BS-P) | 1.130E−1 | 2180.73 |
| Class 3 | CMDE | 2.437 | 653.77 |
| | Kernel with $\psi$ sample (Kern-2) | 9.116E−1 | 18.04 |
| | IWMDE | 2.382 | 638.88 |

Class 2 algorithms. Note that in this example, the likelihood of $\psi$ conditional on $\lambda$ takes the form of normal density, hence the CMDE is available in this case. As one of the features of this example is the restricted integration range of $\lambda$, it has some effect on the implementation of the various algorithms. For example, the high density point $\lambda^*$ mentioned in previous sections can be taken as $\hat{\lambda}_\psi$ in some cases, but in some cases $\hat{\lambda}_{psi}$ will be outside of the integration range. Hence, we could either choose $\lambda^*(j) = max\{(\hat{\psi} - \psi)^2, \hat{\lambda}_\psi\}$ for each dimension $j$, or find the posterior mode and take it as $\lambda^*$.

Consider the breaking load of yarn data given in Cox and Snell (1981, Example Q). There are six bobbins and four observations for each bobbin. Suppose the breaking load of yarn from bobbin $j$ follows a normal distribution with mean $\psi$ and variance $\lambda_j$. Then the integrated likelihood given above can be used to make inference about the common mean $\mu$. Here $m = 6$ and $N_j = 4$ for all $j$.

Given the data, the maximum likelihood estimate of $\psi$ is $\hat{\psi} = 15.505$. To calculate the integrated likelihood, we choose a sequence of $\psi$ from 14 to 17 with a constant increment of 0.05, which means there are 61 points in total. As previous example, the different estimators use 10,000 simulation samples and we replicate the procedure for 100 times. The simulation results are shown in Table 2 and Fig. 2.

The results show that the Chib–Jeliazkov method performs the best overall, followed by the two Bridge Sampling methods. But also note that these three algorithms also cost relatively large amount of time to compute. In this example, these two Bridge Sampling methods are different at the 16th most significant digit. As indicated by Chen (2005), the CMDE estimator and the IWMDE estimator are better than the kernel estimator in general. But in our case, the kernel estimator actually performs a little better. However, it shows that the IWMDE estimator is a good alternative as the CMDE estimator when the CMDE estimator is not available (their performances are quite close). Also, the Laplace estimator is not as acurate as Example 1. The reasons for the difference in the simulation results between Example 1 and Example 2 may be

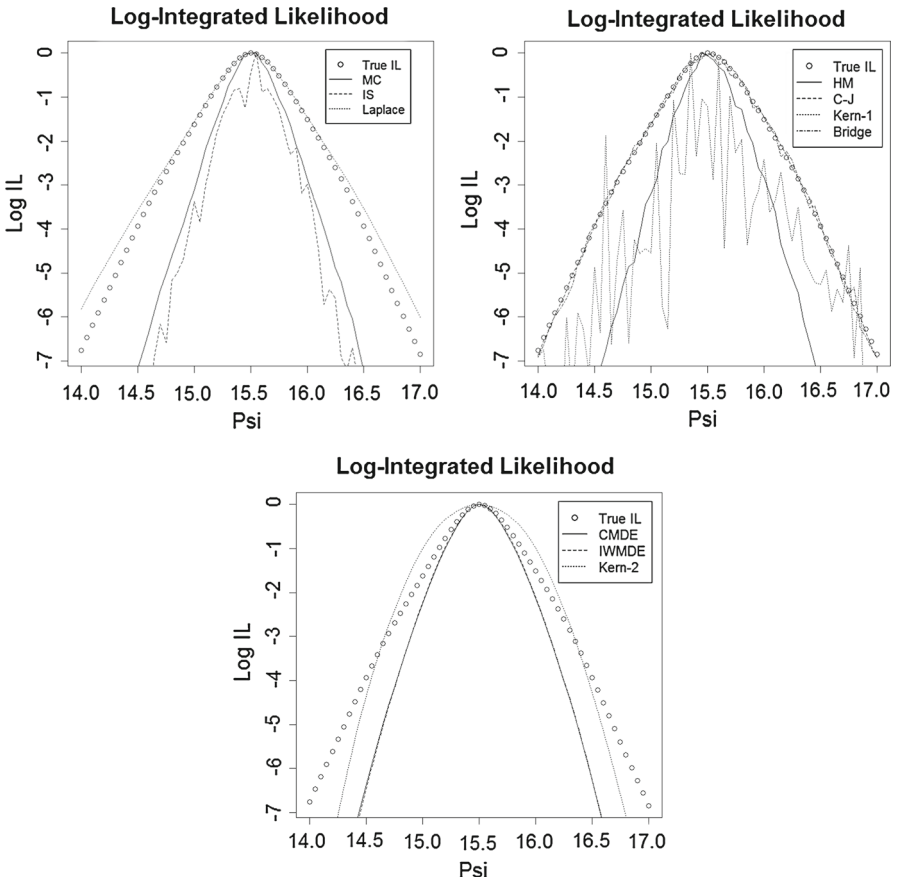


**Fig. 2** Example 2 Class 1, Class 2, and Class 3 estimators

due to the two features of this example: First, the dimension of $\lambda$ is 6 in Example 2 ; Second, the integration range of $\lambda$ depends on $\psi$ in Example 2.

### 4.3 Example 3: Gamma distribution with common shape parameter

Let $Y_{ij}, i = 1, \ldots, q, \ j = 1, \ldots, m$, be independent gamma random variables with shape parameter $\psi$ and scale parameter $1/\lambda_i$, as in Severini (2007). The likelihood is

$$L(\psi, \lambda) = \Pi_{i=1}^{q} \Pi_{j=1}^{m} \frac{Y_{ij}^{\psi-1} \lambda_i^{\psi}}{\Gamma(\psi)} e^{-\lambda_i Y_{ij}}.$$

If the weight function for $\lambda_i$ is taken as the density function of an exponential random variable with mean 1, then the integrated likelihood can be written as

$$\bar{L}(\psi) = \int_{\Lambda} L(\psi, \lambda)\pi(\lambda|\psi)d\lambda$$
$$= \Pi_{i=1}^{q} \int_{-\infty}^{\infty} \Pi_{j=1}^{m} \frac{Y_{ij}^{\psi-1}}{\Gamma(\psi)} \lambda_i^{\psi} e^{-\lambda_i Y_{ij}} e^{-\lambda_i} d\lambda_i.$$

This integral has a closed form and the log integrated likelihood is

$$\bar{l}(\psi) = -mq \log(\Gamma(\psi)) + (\psi - 1) \sum_{i=1}^{q} \sum_{j=1}^{m} log(Y_{ij})$$
$$+ m \log(\Gamma(n\psi + 1)) - \sum_{i=1}^{q} (n\psi + 1) \log \left( \sum_{j=1}^{m} Y_{ij} + 1 \right).$$

The algorithms used for this example are implemented in a similar way to that used in Example 2. For example, we use the uniform distribution as the proposal distribution for the Metropolis–Hastings algorithm. For the Chib method using Gibbs sampler, at $t$th step, a latent variable $z_{it}$ is generated as a gamma random variable with shape parameter $\psi$ and scale parameter $1/\lambda_{i(t-1)}$, where $\lambda_{i(t-1)}$ is the sample point of $\lambda_i$ at step of $t-1$. On the other hand, conditional on $\psi$, $Y$ and $z_t$, the sample point of $\lambda_i$ at step of $t$ is generated following a gamma distribution with shape $n\psi + 1$ and scale $1/(\sum_{j=1}^{m} Y_{ij} + z_t + 1)$. For Class 3 estimators, to generate the sample of $(\psi, \lambda)$, we use the Metropolis–Hastings algorithm with a uniform distribution proposal for both parameters.

For the simulation experiment, we take $q = 2$, $m = 4$ and parameter values $\psi_0 = 2$, $\lambda_0 = (5, 10)$ to generate one sample $\{Y_{ij}\}_{i=1,j=1}^{2,4}$ which is used to compare the different estimators. The integrated likelihood is evaluated at 40 different $\psi$ values equally spaced from 0 to 2. As before, the simulation sample size is 10,000 for all estimators which use simulation. The same procedure is replicated for 100 times using the same $\{Y_{ij}\}_{i=1,j=1}^{2,4}$.

The results are summarized in Table 3 and Fig. 3. Chib–Jeliazkov Method has the smallest RMSE in this example, while bridge sampling method also performs quite well. The Laplace estimator has a poor performance due to the skewed shape of the likelihood function.

## 5 Beta-binomial example on NBA data

In this example, we use the integrated likelihood method on Beta-Binomial model to estimate the average player free-throw percentage for 2013–2014 National Basketball Association (NBA) season. In this data set, there are records of number of attempts and goals made for each of 482 NBA players in this season. The number of attempts of player range from 0 to 1685 with mean 424, and the number of goals made range from 0 to 825 with mean 193. The data set is available at sports.yahoo.com. Raftery Adrian (2006) used beta-binoimal distribution to model the data in 1998–1999 NBA season.

**Table 3** Empirical comparison of different computation methods for example 3

| Class | Method | Mean RMSE | CPU time (seconds per trial) |
|---|---|---|---|
| Class 1 | Monte Carlo (MC) | 1.437E−1 | 10.91 |
| | Importance sampling (IS) | 1.612E−1 | 159.53 |
| | Laplace's method (laplace) | 4.730 | 0.005 |
| Class 2 | Harmonic mean (HM) | 7.691E−1 | 67.58 |
| | Chib method (Chib) | 1.828E−1 | 33.35 |
| | Chib–Jeliazkov method (C-J) | 2.713E−2 | 161.31 |
| | Kernel with $\lambda$ sample (Kern-1) | 1.070 | 75.97 |
| | Bridge-marginal likelihood (BS-M) | 3.838E−2 | 172.22 |
| | Bridge-conditional posterior (BS-P) | 4.142E−2 | 224.75 |
| Class 3 | Kernel with $\psi$ sample (Kern-2) | 2.246 | 6.96 |
| | IWMDE | 1.588E−1 | 219.68 |

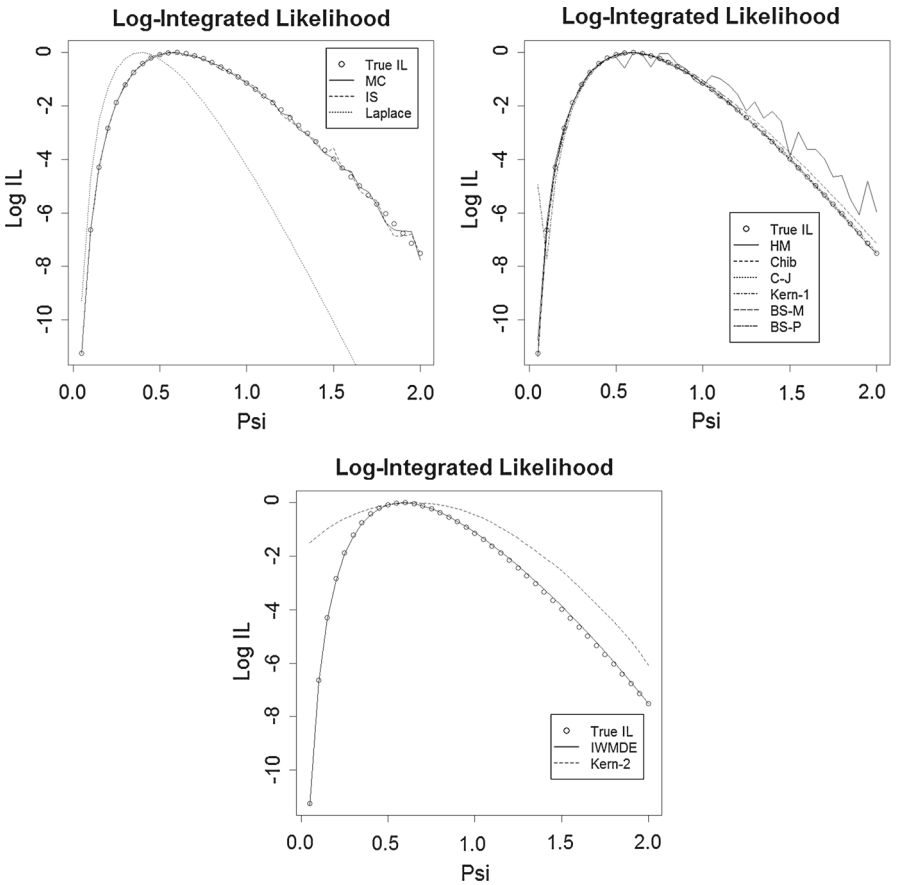


**Fig. 3** Example 3 Class 1, Class 2, and Class 3 estimators

For the $i$th ($i = 1, \ldots, 482$) player, let $X_i$ denote the number of goals made out of total number of attempts $N_i$ with success probability $p_i$. Conditional on $p_i$, assume $X_i$ follows a binomial distribution $Bin(N_i, p_i)$ with probability function:

$$P(X_i|N_i, p_i) = \binom{N_i}{X_i} p_i^{X_i}(1 - p_i)^{N_i - X_i}. \tag{11}$$

The success probability $p_i$ varies from player to player following a beta distribution $Beta(\alpha, \beta)$ with probability density function:

$$p(p_i|\alpha, \beta) = \frac{p_i^{\alpha-1}(1 - p_i)^{\beta-1}}{B(\alpha, \beta)} \tag{12}$$

where $B(\cdot, \cdot)$ denotes beta function.

Combining (11) and (12), $X_i$ follows a Beta-Binomial distribution $Beta - Bin(\alpha, \beta)$ with probability function:

$$P(X_i|N_i, \alpha, \beta) = \int_0^1 P(X_i|N_i, p_i)p(p_i|\alpha, \beta)dp_i \tag{13}$$

$$= \binom{N_i}{X_i} \frac{B(X_i + \alpha, N_i - X_i + \beta)}{B(\alpha, \beta)}. \tag{14}$$

Under the Beta-Binomial distribution assumption, the expectation of average free-throw percentage is $E[X_i/N_i|N_i, \alpha, \beta] = \frac{\alpha}{\alpha+\beta}$. To estimate the average free-throw percentage, a more convenient parameterization is to let $\psi = \frac{\alpha}{\alpha+\beta}$ and $\lambda = \frac{1}{\alpha+\beta+1}$, then $E(X_i) = N\psi$, and $V(X_i) = N\psi(1 - \psi)[1 + (n - 1)\lambda]$. The parameter of interest, i.e. average free-throw percentage, is $\psi$ and the nuisance parameter is $\lambda$. Both parameters have support range from 0 to 1. Probability function (13) can be rewritten as

$$P(X_i|N_i, \psi, \lambda) = \binom{N_i}{X_i} \frac{B(X_i + \psi(1/\lambda - 1), N_i - X_i + (1 - \psi)(1/\lambda - 1))}{B(\psi(1/\lambda - 1), (1 - \psi)(1/\lambda - 1))}.$$

The integrated likelihood method can be used to estimate the average free-throw percentage parameter $\psi$. As the closed form of integrated likelihood is unavailable, numerical computation algorithms are useful in this scenario. One method from each class is chosen, and they are importance sampling method, Chib–Jeliazkov method, and IWMDE method respectively. Three different prior functions are used to evaluate empirical prior effect on the integrated likelihood function and computation methods. Simulation sample sizes used are 10,000. The results are summarized in Fig. 4.

Among these three methods, importance sampling seems to be the least stable one. The lack of smoothness is a sign of inaccuracy. There is a small but persistent difference between Chib–Jeliazkov method and IWMDE method under prior Beta (1, 1), especially on the left side of the mode. Though we don't have closed form of integrated likelihood as benchmark, looking at the three curves, importance sampling is closer to Chib–Jeliazkov method than to IWMDE, that can be an indication that Chib–

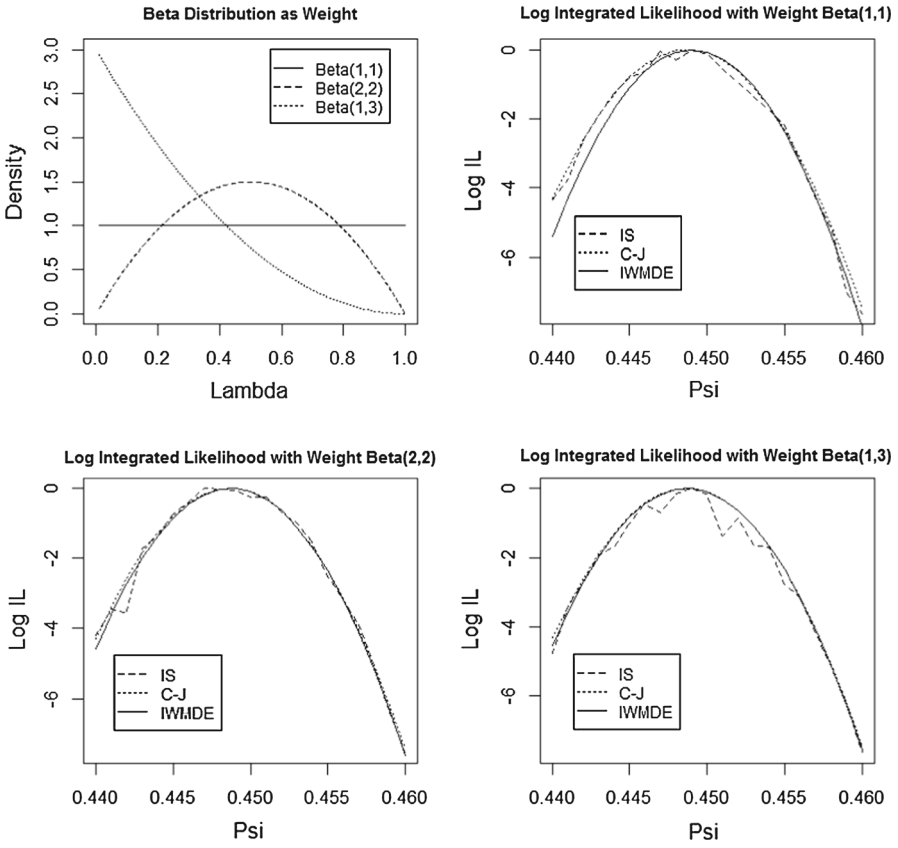


**Fig. 4** Beta-binomial example on NBA data with 3 prior functions

Jeliazkov is closer to the true integrated likelihood than IWMDE. With these evidences, Chib–Jeliazkov seems to be the most reliable estimator in this example. We examined three different priors in this example. There is little effect of priors on integrated likelihood function itself. The priors affect the performance of the algorithms in some degree, but not very notable. For example, the IWMDE performs a little better under Beta(2,2) and Beta(1,3) than Beta(1,1); Importance Sampling has the largest variation with Beta(1,3); Chib–Jeliazkov performs consistently well with three priors.

The average (w.r.t. different weight functions) CPU time is 52.90 s for Importance Sampling method, 263.73 s for Chib–Jeliazkov method, and 147.30 s for IWMDE method. The time is counted for computing integrated likelihood at $20\psi$ points in the interval of [0.440, 0.460], on a machine with Intel Core i7-2760QM 2.40GHz CPU and 8GB memory. The relative time complexity is similar to previous examples.

## 6 Discussion

This paper discusses 12 methods for computing an integrated likelihood function and compares these methods by applying them to both simulation and real data analysis

**Table 4** Summary of computation methods in accuracy, computation time, and implementation effort

| Class | Method | Accuracy | Computation time | Implementation effort |
|---|---|---|---|---|
| Class 1 | Monte Carlo (MC) | * | *** | *** |
| | Importance sampling (IS) | * | *** | *** |
| | Laplace's method (laplace) | *** (Simple) | *** | *** |
| | | * (Complex) | | |
| Class 2 | Harmonic mean (HM) | * | ** | ** |
| | Chib method (Chib) | ** | ** | * |
| | Chib–Jeliazkov method (C-J) | *** | * | * |
| | Bridge-Marginal likelihood (BS-M) | *** | * | * |
| | Bridge-Conditional posterior (BS-P) | *** | * | * |
| | Kernel with $\lambda$ sample (Kern-1) | * | ** | ** |
| Class 3 | Kernel with $\psi$ sample (Kern-2) | * | *** | ** |
| | IWMDE | ** (Simple) | ** | * |
| | | * (Complex) | | |
| | CMDE | ** (Simple) | ** | * |
| | | * (Complex) | | |

More stars means better: higher accuracy, faster computation, and lower effort. The accuracy of some methods vary in different scenarios, and it is differentiated by accuracy ratings for simple models and accuracy ratings for complex models

examples. The results are summarized in Table 4. These methods are evaluated in accuracy, computation time (performance), and implementation effort. The Chib–Jeliazkov method is consistently shown as accurate and robust for different scenarios, though it has higher cost in computation time and implementation effort due to MCMC sampling procedure. The properties of the two bridge sampling estimators are very similar to Chib–Jeliazkov method. The Chib method can provide a good estimate and takes a relatively short computing time, but it usually requires extra effort to apply due to the data augmentation and knowledge of conditional distribution. The two kernel estimators are not accurate and will suffer the curse of dimensionality when the dimension of the parameter is high. IWMDE produces a smooth estimation of the integrated likelihood function, however, the accuracy varies from case to case. For example, when $\hat{\lambda}_\psi$ is far away from $\hat{\lambda}$, it will produce a poor estimate of the integrated likelihood. The direct approaches (Class 1) are not robust in accuracy across different cases, though they are less time consuming in general. For instance, the Laplace estimator can be extremely accurate in some cases when the underlying likelihood has a symmetric shape with the mode in the middle, but when the shape is skewed, the accuracy may be poor. The empirical evidence shows the performance of these 12 methods are affected by the weight (prior) function in some degree, but not remarkable. Overall, the Chib–Jeliazkov method and the two bridge sampling methods perform well in the empirical evaluation and, based on these results, we recommend these methods for computation of the integrated likelihood.

**Acknowledgments** The work of T. A. Severini was supported by NSF Grant DMS-1308009. This research was supported in part through the computational resources and staff contributions provided for the Social Sciences Computing cluster (SSCC) at Northwestern University. Recurring funding for the SSCC is provided by Office of the President, Weinberg College of Arts and Sciences, Kellogg School of Management, the School of Professional Studies, and Northwestern University Information Technology.

## References

Berger JO, Liseo B, Wolpert RL (1999) Integrated likelihood methods for eliminating nuisance parameters. Stat Sci 14(1):1–22

Barndorff-Nielsen O (1980) Conditionality resolutions. Biometrika 67(2):293–310

Barndorff-Nielsen O (1983) On a formula for the distribution of the maximum likelihood estimator. Biometrika 70(2):343–365

Barndorff-Nielsen OE (1995) Stable and invariant adjusted profile likelihood and directed likelihood for curved exponential models. Biometrika 82(3):489–499

Barndorff-Nielsen OE, Nielsen OEB (1989) Asymptotic techniques; for use in statistics

Bos CS (2002) A comparison of marginal likelihood computation methods. In: COMPSTAT 2002—proceedings in computational statistics, pp 111–117

Chen M-H (1994) Importance-weighted marginal bayesian posterior density estimation. J Am Stat Assoc 89:818–824

Chen M-H (2005) Bayesian computation: from posterior densities to Bayes Factors, marginal likelihoods, and posterior model probabilities. Handb Stat 25:437–457

Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90:1313–1321

Chib S, Greenberg E (1995) Understanding the metropolis hastings algorithm. Am Stat 49:327–335

Chib S, Jeliazkov I (2001) Marginal likelihood from the Metropolis–Hastings output. J Am Stat Assoc 96:270–281

Cox CP (1985) Interval estimates for the ratio of the means of two normal populations with variances related to the means. Biometrics 41:261–265

Cox DR, Snell EJ (1981) Applied statistics: principles and examples. Chapman and Hall, London

Gelfand AE, Smith AFM, Lee TM (1992) Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. J Am Stat Assoc 87:523–532

Givens GH, Hoeting JA (2005) Computational statistics. Wiley, New York

Hastings WK (1970) Monte Carlo sampling methods using Markov Chains and their applications. Biometrika 57:97–109

Jensen JL (1995) Saddle point approximations. Oxford University Press, Oxford

Kalbfleisch JD, Sprott DA (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). J R Stat Soc B 32:175–208

Kass RE, Tierney L, Kadane JB (1990) Bayesian and likelihood methods in statistics and econometrics: essays in honor of George A, Barnard. Amsterdam, pp 473–488

Kass RE, Carlin BP, Gelman A, Neal RM (1998) Markov Chain Monte Carlo in practice: a roundtable discussion. Am Stat 52(2):93–100

Li Q, Racine JS (2007) Nonparametric econometrics: theory and practice. Princeton University Press, Princeton

Liseo B (1993) Elimination of nuisance parameters with reference priors. Biometrika 80:295–304

Mendoza M, ITAM RH, Angel S (2005) Inferences on the ratio of normal means and other related problems. Estadistica 57:168–169

Meng XL, Schilling S (2002) Warp Bridge sampling. J Comput Graph Stat 11(3):552–586

Meng XL, Wong WH (1996) Simulating ratios of Normalizing constants via a simple identity: a theoritical exploration. Stat Sin 6:831–860

Mira A, Nicholls G (2004) Bridge estimation of the probability density at a point. Stat Sin 14:603–612

Newton MA, Raftery AE (1994) Approximate Bayesian inference by the weighted likelihood bootstrap. J R Stat Soc Ser B 3:3–48

Raftery AE, Newton MA, Satagopan JM, Krivitsky PN (2006) Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. bepress

Severini TA (2005) Integrated likelihood functions for non-Bayesian inference. (Full version)

Severini TA (1998) An approximation to the modified profile likelihood function. Biometrika 85(2):403–411

Severini TA (1999) On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. Stat Sin 9:713–724
Severini TA (2000) Likelihood methods in statistics. Oxford University Press, Oxford
Severini TA (2007) Integrated likelihood functions for non-Bayesian inference. Biometrika 94:529–542
Severini TA (2010) Likelihood ratio statistics based on an integrated likelihood. Biometrika 97:481–496
Severini TA (2011) Frequency properties of inferences based on an integrated likelihood. Stat Sin 21:433–447
Silverman BW (1986) Density estimation for statistics and data analysis. CRC Press, Boca Raton
Srivastava MS (1986) Multivariate bioassay, combination of bioassays and Fieller's theorem. Biometrics 42:131–141
Tanner MA (1996) Tools for statistical inference: methods for the exploration of posterior deistributions and likelihood functions, 3rd edn. Springer, Berlin
Tanner MA, Wong W (1987) The calculation of posterior distributions by data augmentation (with discussion). J Am Stat Assoc 82:528–550
Tierney L, Kadane JB (1986) Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc 81(393):82–86
Wei GC, Tanner MA (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J Am Stat Assoc 85(411):699–704