Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2013-42

Importance Sampling Schemes for Evidence Approximation in Mixture Models

J. E. LEE[1]
C. P. ROBERT[2]

December 2013

[1] Auckland University of Technology, New Zealand. Email : jelee@aut.ac.nz.com

[2] PSL, Université Paris-Dauphine, CEREMADE ; CREST, PARIS and University of Warwick, Department of Statistics. Email : xian@ceremade.dauphine.fr

Importance sampling schemes for evidence approximation in mixture models

Jeong Eun Lee * and Christian, P. Robert †

Abstract.

The marginal likelihood is a central tool for drawing Bayesian inference about the number of components in mixture models. It is often approximated since the exact form is unavailable. A bias in the approximation may be due to an incomplete exploration by a simulated Markov chain (e.g., a Gibbs sequence) of the collection of posterior modes, a phenomenon also known as lack of label switching, as all possible label permutations must be simulated by a chain in order to converge and hence overcome the bias. In an importance sampling approach, imposing label switching to the importance function results in an exponential increase of the computational cost with the number of components. In this paper, two importance sampling schemes are proposed through choices for the importance function; a MLE proposal and a Rao-Blackwellised importance function. The second scheme is called dual importance sampling. We demonstrate that this dual importance sampling is a valid estimator of the evidence and moreover show that the statistical efficiency of estimates increases. To reduce the induced high demand in computation, the original importance function is approximated but a suitable approximation can produce an estimate with the same precision and with reduced computational workload.

Keywords: Model evidence, Importance sampling, Mixture models, Marginal likelihood

1 Introduction

We consider $x = (x_1, \cdots, x_n)$, a sample of (univariate or multivariate) observations from a finite mixture of k distributions,

$$x \sim f_k(x|\theta) = \sum_{i=1}^{k} \lambda_i f(x|\xi_i)\,, \qquad \sum_{i=1}^{k} \lambda_i = 1\,.$$

The number of components, k, is often a (or even *the*) quantity of interest. Following the perspective on mixtures adopted in Richardson and Green (1997), our inference about k is based on the evidence or marginal likelihood $\mathfrak{E}(k)$,

$$\mathfrak{E}(k) = \int_S f_k(x|\theta)\pi_k(\theta)\,\mathrm{d}\theta$$

*Auckland University of Technology, New Zealand jelee@aut.ac.nz.com

†PSL, Université Paris-Dauphine, CEREMADE, Department of Statistics, University of Warwick, and CREST, Paris xian@ceremade.dauphine.fr

where S is the state space for θ. The ratio of evidences is a Bayes factor and is properly scaled to be readily compared to 1 (Jeffreys 1939). The difficulty with this approach is that the quantities $\mathfrak{E}(k)$ are usually unavailable and cannot be directly and reliably (Newton and Raftery 1994) derived from simulations from the posterior distribution on θ. In such cases, the marginal likelihood is approximated using a dedicated method and this approximation is used for a model comparison with respect to k.

Taking advantage of the traditional missing data structure of mixture models, we introduce missing (or dummy) variables $z = (z_1, \ldots, z_n)$ such that $(j = 1, \ldots, n)$

$$x_j | z_j \sim f(x_j | \xi_{z_j}) \,,$$

and

$$z_j | x_j \sim \mathcal{M}\left(\frac{\lambda_1 f(x_j|\xi_1)}{\sum_{i=1}^k \lambda_i f(x_j|\xi_i)}, \ldots, \frac{\lambda_k f(x_j|\xi_k)}{\sum_{i=1}^k \lambda_i f(x_j|\xi_i)} \right) .$$

This representation of the mixture distribution allows the data being clustered and each cluster brings information about a parameter ξ_j for the corresponding component j. In particular, when the full conditional distribution $f(x_j|\xi_{z_j})$ is available in closed form, conditional simulation from $\pi(\zeta_1, \ldots, \zeta_k | \mathbf{x}, \mathbf{z})$ becomes easier, as demonstrated in Diebolt and Robert (1994), in connection (Dempster et al. 1997) with the expectation maximisation (EM) algorithm.

The drawback of this scheme (which leads to the Gibbs sampler) is that it considerably slows down the exploration abilities of the algorithm, most usually causing the resulting Markov chain to get stuck in a certain region of the highly multimodal posterior distribution. When this Markov chain gets trapped around one mode, it obviously suffers from a certain type of lack of convergence. On the other hand, point estimates of the component-wise parameters are harder to construct when the Markov chain moves freely between modes. This phenomenon is known under the name of label switching (Celeux et al. 2000; Frühwirth-Schnatter 2001; Jasra et al. 2005) or lack thereof. For the issue of numerically approximating the evidence, a lack of exploration over all significant modes has a clear impact on the numerical accuracy of the approximation, as exhibited most forcibly by Neal (1999).

In this paper, we focus on numerical marginal likelihood approximations using importance sampling (IS) and investigate suitable choices for importance functions, taking into consideration the label-switching requirement of exploring all $k!$ replicas of the posterior modes. Two approaches are proposed; (i) Chib's version of importance sampling and (ii) a novel representation called *dual importance sampling*. For computational reasons, an approximate version of the dual importance sampling algorithm is investigated towards an increase in computational efficiency, which is obtained by avoiding negligible function evaluations for the proposal density q. Those approaches are compared with standard estimators using Chib's approach and bridge sampling. The performances are evaluated on simulated and benchmark datasets, namely the galaxy and fishery datasets

used in Richardson and Green (1997).

The paper starts with a brief summary of the approximations in Chib's method and bridge sampling in Section 1. In Section 2, importance sampling is studied and the choices considered for the importance function are described. Our importance function approximate approach is introduced in Section 3. Simulation studies using simulated and real datasets are reported in Section 4, and the paper concludes with a short discussion in Section 5.

2 Standard evidence estimators

2.1 Chib's estimator

The reference estimator for evidence approximation is Chib's (1995) representation of the marginal likelihood

$$m_k(x) = \frac{\pi_k(\theta^o) f_k(x|\theta^o)}{\pi_k(\theta^o|x)}$$

which holds for *any* choice of the plug-in value θ^o. (It was also called *the candidate's formula* by the late Julian Besag.) While $\pi_k(\theta^o|x)$ is not available in closed form for mixtures, the Gibbs sampling decomposition allows for a Rao–Blackwellised approximation (Robert and Casella 2004) that converges at a parametric speed, as already noticed in Gelfand and Smith (1990):

$$\hat{\pi}_k(\theta^o|x) = \frac{1}{T}\sum_{t=1}^{T} \pi_k(\theta^o|x, z^t)\,.$$

However, convergence of this estimate is guaranteed only when label switching has occurred within the simulated Markov chain. When $(z_1^t, \ldots, z_n^t)$ remains instead concentrated around a single mode, the approximation of $\log m_k(x_1, \ldots, x_n)$ using Chib's representation is usually off by a factor of order $O(\log k!)$, even though this later term cannot be used as a reliable correction as noted by Neal (1999).

To deal with the resulting bias, thus induced by the limited exploration provided by a non-label-switching Markov chain, Berkhof et al. (2003) suggested a straightforward correction in which another Rao-Blackwellisation is undertaken, namely the replacement of the above by an average over all possible permutations of the labels, forcing the label switching and the exchangeability of the labels to occur in a "perfect" manner: the resulting approximation can be expressed as

$$\tilde{\pi}_k(\theta^o|x) = \frac{1}{Tk!}\sum_{\sigma\in\mathfrak{S}(k)}\sum_{t=1}^{T} \pi_k(\theta^o k|x, \sigma(z^t))\,,$$

where $\mathfrak{S}(k)$ denotes the set of the $k!$ permutations of $\{1, \ldots, k\}$ and where σ is one of

those permutations. Note that the above correction can also be rewritten as

$$\tilde{\pi}_k(\theta^o|x) = \frac{1}{Tk!} \sum_{\sigma \in \mathfrak{S}(k)} \sum_{t=1}^{T} \pi_k(\sigma(\theta^o)|x, z^t)\,, \tag{1}$$

which may induce some computational savings.

While Chib's representation has often been advocated as a reference solution for computing the evidence, other solutions abound within the literature, from nested sampling (Skilling 2007; Chopin and Robert 2010) to reversible jump MCMC (Green 1995; Richardson and Green 1997), to particle filtering (Chopin 2002).

2.2 Bridge Sampling

Bridge sampling is a technique for computing ratios of normalising constants based on iid or MCMC simulations from posterior distributions (Meng and Wong 1996). Mira and Nicholls (2004) used it as a point posterior estimate for Chib's method and it is well-suited to estimate the marginal likelihood (Frühwirth-Schnatter 2001, 2004). The generic equality at the core of bridge sampling is given by

$$\widehat{\mathfrak{E}}(k) = \frac{\mathbb{E}_q(\alpha(\theta)\pi^*(\theta|x))}{\mathbb{E}_\pi(\alpha(\theta)q(\theta))} \tag{2}$$

where $\pi^*(\theta|X)$ is the unnormalised posterior distribution obtained as the product of the prior density and the likelihood function and where α is an arbitrary function such that the expectations are well-defined over the joint support of q and π^* (Chen et al. 2000).

The (formally) optimal choice for the function α by Meng and Wong (1996) leads to the following iterative approximate of the evidence

$$\widehat{\mathfrak{E}}^{(t)}(k) = \widehat{\mathfrak{E}}^{(t-1)}(k)\, \frac{L^{-1}\sum_{l=1}^{L} \dfrac{\hat{\pi}(\tilde{\theta}^l|x)}{Lq(\tilde{\theta}^l) + M\hat{\pi}(\tilde{\theta}^l|x)}}{M^{-1}\sum_{m=1}^{M} \dfrac{q(\theta^m)}{Lq(\theta^m) + M\hat{\pi}(\theta^m|x)}}\,, \tag{3}$$

where $\tilde{\theta}^1, \ldots, \tilde{\theta}^L \sim q$ and $\theta^1, \ldots, \theta^M \sim \pi(\cdot|x)$, possibly by an MCMC algorithm. Here, $\hat{\pi}(\theta|x) = \pi^*(\theta)/\widehat{\mathfrak{E}}^{(t-1)}(k)$.

This optimal α and the resulting bridge sampling are trivial to implement and to check for convergence when π^* and q have overlapping support. If the intersection of the supports is too small, the resulting bridge sampling estimate may turn out to be too variable (Voter 1985; Servidea 2002). For such cases, methods such as path sampling (Gelman and Meng 1998), a simple location shift of a distribution to the other distribution (Voter 1985), and a warp bridge sampling (Meng and Schilling 2002) have

been proposed to reduce a difference between supports of two distributions in terms of shape and dimensions.

For the choice of q, Frühwirth-Schnatter (2001) proposed a Rao-Blackwellised importance function of the form

$$q(\theta) = \frac{1}{J}\sum_{j=1}^{J} \pi_k(\theta|\theta^{(j)}, z^{(j)}, x) \tag{4}$$

where $\{\theta^{(j)}, z^{(j)}\}_{j=1}^{J}$ are well-mixed simulations from the posterior, thus imposing the label switching. This form is such that its support properly overlaps with the support of π^* and the bridge sampling (3) is applicable. Frühwirth-Schnatter (2004) demonstrated that this iterative estimate converges relatively quickly, in about $T = 10$ iterations, even with different starting values.

3 Further importance sampling estimators

3.1 A plug-in proposal

Recall that the importance weight $\omega(\theta)$ in the importance sampling algorithm with proposal q is given by

$$\omega(\theta) = \pi_k(\theta) f_k(x|\theta) \big/ q(\theta)\,, \tag{5}$$

and that it leads to an evidence approximation of the form

$$\widehat{\mathfrak{E}}(k) = \frac{1}{T}\sum_{t=1} \omega(\theta^t)\,. \tag{6}$$

Using a Rao-Blackwell argument inspired from Chib's representation, a natural importance function is $q(\theta) = \pi_k(\theta|x, z^o)$, which generates samples from the posterior conditional on a completion vector z^o, for instance the MAP or the marginal MAP estimate of $\mathbf{z}$ derived from an MCMC run.

While this estimator is theoretically valid, providing an unbiased estimator of $\widehat{\mathfrak{E}}(k)$, it may face difficulties in practice by missing wide regions of the parameter space when simulating from $\pi_k(\theta|x, z^o)$. This is indeed the practical version of simulating from an importance function with a support that is smaller than the support of the integrand, a setting that leads to an erroneous approximation of the corresponding integral. In the current situation, since $\pi_k(\theta|x, z^o)$ is everywhere positive, this is not a theoretical issue. However, in practice, the conditional density is numerically equal to zero around the alternative modes.

Inspired by Berkhof et al. (2003), we thus propose to replace the MAP proposal by

its symmetrised version

$$q(\theta) = \frac{1}{k!} \sum_{\sigma \in \mathfrak{S}(k)} \pi_k(\sigma(\theta)|x, z^o)\,. \tag{7}$$

with a natural notational shortcut using σ as the permutation applying to the indices of different components of θ. This new proposal is equivalent to generating θ from the original conditional distribution and then operating a random permutation on the components of θ. The computational cost of producing $\omega(\theta)$, hence $\widehat{\mathfrak{E}}(k)$, is then multiplied by $k!$ but the support of the proposal hopefully gets wide enough to include all the modes of the target distribution.

If the tails of this proposal are deemed to be too narrow (as signalled by the effective sample size), additional values of $z^t = (z_1^t, \ldots, z_n^t)$ can be extracted from the Gibbs chain to robustify the proposal, provided the symmetry is recovered by means of the same averaging over all permutations.

3.2 Dual importance sampling

A dual exploitation of the Rao-Blackwellisation argument produces an alternative importance sampling proposal, based on MCMC draws of (θ, z) conditional on $x = (x_1, \ldots, x_n)$ and the "current values" $(\theta^{(j)}, z^{(j)})$ $(j = 1, \ldots, J)$,

$$\begin{aligned} q(\theta) &= \frac{1}{Jk!} \sum_{j=1}^{J} \sum_{\sigma \in \mathfrak{S}(k)} \pi_k(\sigma(\theta)|\theta^{(j)}, z^{(j)}, x) \\ &= \frac{1}{Jk!} \sum_{j=1}^{J} \sum_{\sigma \in \mathfrak{S}(k)} \pi_k(\theta|\sigma(\theta^{(j)}, z^{(j)}), x)\,. \end{aligned} \tag{8}$$

Here, $\pi_k(\theta|\sigma(\theta^{(j)}, z^{(j)}), x)$ denotes a product of conditional densities on each (or subset) of unknown parameters $\theta = (\lambda_1, \ldots, \lambda_k, \xi_1, \ldots, \xi_k)$ in a Gibbs sampler representation. The label switching is imposed upon those conditional densities in all $k!$ ways and an average of J conditional densities approximates any one of $k!$ symmetric terms of a marginal posterior with the same weight.

Both formulas (4) by Frühwirth-Schnatter (2001) and (8) have the same underlying motivation of approximating a multimodal marginal posterior. For (4), the label switching is already occurred in selected conditional densities for a Monte Carlo approximation. For the later one, a marginal posterior is approximated in two ways, a Monte Carlo approximation for one of $k!$ symmetric density terms and $k!$ permutations of $\{\sigma(\theta^{(j)}, z^{(j)})\}_{j=1}^{J}$ for the multimodality. As k increases, a number of conditional densities for (4) increases exponentially to approximate all $k!$ terms while it is not necessary

for (8). Those two forms induce different efficiency properties for marginal likelihood estimates.

It is well-known that the posterior distribution contains $k!$ symmetric terms. Without loss of generality, a posterior sample θ is represented by two components, ν and κ such that $\theta = \sigma_\kappa(\nu)$. Here, κ denotes a term allocation or index, $\kappa \in \{1, \ldots, k!\}$, and ν is a random variable generated from any term of the posterior density. Using a common prior for all components, the posterior densities for ν in all possible permutation transforms are equal:

$$\pi(\sigma_1(\nu))f_k(x|\sigma_1(\nu)) = \cdots = \pi(\sigma_{k!}(\nu))f_k(x|\sigma_{k!}(\nu)) \ .$$

A marginal likelihood estimate using $q(\nu, \kappa|x)$ is based upon the traditional importance sampling identity

$$\mathfrak{E}(k) = \sum_{\kappa=1}^{k!} \int \frac{\pi(\nu,\kappa)f_k(x|\nu,\kappa)}{q(\nu,\kappa|x)} q(\nu,\kappa|x)\mathrm{d}\nu = \mathbb{E}_{q(\nu,\kappa|x)}[\omega(\nu,\kappa)]$$

leading to

$$\hat{\mathfrak{E}}(k) = \frac{1}{T}\sum_{t=1}^{T} \omega(\nu^{(t)}, \kappa^{(t)}) \tag{9}$$

where $\omega(\nu,\kappa) = \pi(\nu,\kappa)f_k(x|\nu,\kappa)/q(\nu,\kappa|x)$. Rewriting $q(\nu,\kappa|x) = p(\nu|x)p(\kappa|x)$ and marginalizing $q(\nu,\kappa|x)$ over κ, the estimate of

$$\mathfrak{E}(k) = \int \left(\sum_{\kappa=1}^{k!} \frac{\pi(\nu,\kappa)f_k(x|\nu,\kappa)}{q(\nu,\kappa|x)} p(\kappa|x) \right) p(\nu|x) \ \mathrm{d}\nu = \mathbb{E}_{p(\nu|x)}[\omega(\nu)]$$

becomes

$$\hat{\mathfrak{E}}(k) = \frac{1}{T}\sum_{t=1}^{T} \omega(\nu^{(t)}) \ , \qquad \nu^{(t)} \sim p(\nu|x) \tag{10}$$

where $\omega(\nu) = \mathbb{E}_{p(\nu|x)}[\pi(\nu,\kappa)f_k(x|\nu,\kappa)/q(\nu,\kappa|x)]$.

The estimate (9) is equivalent to (4) in which q is constructed using randomly permuted Gibbs samples. In the importance function (8) the J-size Gibbs samples are permuted in $k!$ ways with equal probability. When $p(\kappa|x) = 1/k!$, the estimate (10) is equivalent to the dual importance sampling version. By the law of large numbers, both estimates (9) and (10) converge to $\mathfrak{E}(k)$,

$$\frac{1}{T}\sum_{t=1}^{T} \omega(\nu^{(t)}, \kappa^{(t)}) \overset{T\to\infty}{\underset{p}{\longrightarrow}} \mathfrak{E}(k) \qquad \text{and} \qquad \frac{1}{T}\sum_{t=1}^{T} \omega(\nu^{(t)}) \overset{T\to\infty}{\underset{p}{\longrightarrow}} \mathfrak{E}(k) \ .$$

Furthermore, a Central Limit theorem holds,

$$\sqrt{T}\left\{\frac{1}{T}\sum_{t=1}^{T}\omega(\nu^{(t)},\kappa^{(t)})-\mathfrak{E}(k)\right\}\underset{T\to\infty}{\longrightarrow}\mathcal{N}(0,V_1)$$

and

$$\sqrt{T}\left\{\frac{1}{T}\sum_{t=1}^{T}\omega(\nu^{(t)})-\mathfrak{E}(k)\right\}\underset{T\to\infty}{\longrightarrow}\mathcal{N}(0,V_2)$$

where $V_1=\mathrm{var}_{q(\nu,\kappa|x)}(\omega(\nu,\kappa))$ and $V_2=\mathrm{var}_{p(\nu|x)}(\omega(\nu))$. By the variance decomposition

$$\begin{array}{rcl}\mathrm{var}_{q(\nu,\kappa|x)}(\omega(\nu,\kappa)) & = & \mathrm{var}_{p(\nu|x)}(\mathbb{E}_{p(\kappa|x)}(\omega(\nu,\kappa)))+\mathbb{E}_{p(\nu|x)}(\mathrm{var}_{p(\kappa|x)}(\omega(\nu,\kappa)))\\ & = & \mathrm{var}_{p(\nu|x)}(\omega(\nu))+\mathbb{E}_{p(\nu|x)}(\mathrm{var}_{p(\kappa|x)}(\omega(\nu,\kappa)))\ .\end{array}$$

It is easy to see that $V_1\geq V_2$ and the asymptotic variance of $\hat{\mathfrak{E}}(k)$ associated with q in (4) is thus greater than the one corresponding to (8).

It is also obvious that the computational workload associated with such q's increases exponentially with k. This computational challenge becomes a severe drawback in practice as Berkhof et al. (2003) and Frühwirth-Schnatter (2004) acknowledged.

4 Importance function approximation

4.1 An alternative invariant representation

Suppose that $\{\varphi^{(j)}\}_{j=1}^{J}$ denotes a set of hyperparameters, that $\{\theta^{(j)},z^{(j)}\}_{j=1}^{J}$ is as in (8) and that there is no label switching (or that any label switching has been removed). The $k!$ permutation acting on φ being $\sigma_1,\ldots,\sigma_{k!}$, equation (8) can be rewritten as

$$q(\theta)=\frac{1}{Jk!}\sum_{j=1}^{J}\sum_{i=1}^{k!}\pi(\theta|\sigma_i(\varphi^{(j)}),x)=\frac{1}{k!}\sum_{i=1}^{k!}h_{\sigma_i}(\theta) \tag{11}$$

where $h_{\sigma_i}(\theta)=\dfrac{1}{J}\displaystyle\sum_{j=1}^{J}\pi(\theta|\sigma_i(\varphi^{(j)}),x)$. Each of the densities $h_{\sigma_1},\cdots,h_{\sigma_{k!}}$ has a specific support denoted by $S_1,\cdots,S_{k!}$ which union is the support of q, namely $\mathbb{S}=\bigcup_{i=1}^{k!}S_i$.

From a computational perspective, an artificial label switching step is necessary in computing q but not generating from it. Indeed, assume that the permutation representations $\sigma_1,\ldots,\sigma_{k!}$ are acting on both θ and φ. For an arbitrary permutation representation $\sigma_c(\theta)$ and $\sigma_i(\varphi)$, the following holds

$$\pi(\sigma_c(\theta)|\sigma_i(\varphi),x)=\pi(\sigma_m\sigma_c(\theta)|\sigma_m\sigma_i(\varphi),x)\ ,\quad\forall\ \sigma_m\in\{\sigma_1,\cdots,\sigma_{k!}\}$$

where $\sigma_m\sigma_c(\theta)$ indicates the composition of two permutations acting on θ. Since the full permutation representation set $\{\sigma_1, \cdots, \sigma_{k!}\}$ is equal to $\{\sigma_m\sigma_1, \cdots, \sigma_m\sigma_{k!}\}$, the q-values for $\sigma_c(\theta)$ and $\sigma_m\sigma_c(\theta)$ are equal,

$$\begin{aligned} q(\sigma_c(\theta)) &= \frac{1}{Jk!}\sum_{j=1}^{J}\sum_{i=1}^{k!}\pi(\sigma_c(\theta)|\sigma_i(\varphi^{(j)}), x) = \frac{1}{Jk!}\sum_{j=1}^{J}\sum_{i=1}^{k!}\pi(\sigma_m\sigma_c(\theta)|\sigma_m\sigma_i(\varphi^{(j)}), x) \\ &= \frac{1}{Jk!}\sum_{j=1}^{J}\sum_{i=1}^{k!}\pi(\sigma_m\sigma_c(\theta)|\sigma_i(\varphi^{(j)}), x) = q(\sigma_m\sigma_c(\theta)) \ . \end{aligned} \tag{12}$$

This allows us to state that, when particles are generated from (11), corresponding q-values are equal whether particles are relabelled or not. This holds even when particles are relabelled by imposing an artificial identifiability constraint and the label switching is removed.

The marginal likelihood estimate based on q in (11) is equivalent (from a computational viewpoint) to one based on particles that are generated without imposing label switching when the importance weights are computed according to (11). We assume that all particles are generated from h_{σ_1}. For each particle, a q computation consists of evaluating all of $h_{\sigma_1}, \ldots, h_{\sigma_{k!}}$. Then, a question of interest is to determine which ones of the $h_{\sigma_1}, \ldots, h_{\sigma_{k!}}$ are likely to be insignificant (almost zero) for any θ generated from $h_{\sigma_1}(\cdot)$. In other words, the issue is in determining the amount of overlap in the permuted components.

The next section explains how to approximate the q-computations for particles generated from a particular h, say h_c, by using only significant functions among $h_{\sigma_1}, \ldots, h_{\sigma_{k!}}$ in (11). This means finding the approximate set $\mathfrak{A}(k) \subseteq \mathfrak{S}(k)$ which contains the permutation representations significantly contributing to q.

4.2 Double importance sampling using an approximation

Both $h_{\sigma_i}(\theta)$ and $q(\theta)$ are the functions of random variable, $\theta \sim h_{\sigma_c}(\theta)$, and the contribution in h_{σ_i} relative to q is expressed as

$$\eta_{\sigma_i}(\theta) = \frac{h_{\sigma_i}(\theta)}{k!q(\theta)} = \frac{h_{\sigma_i}(\theta)}{\sum_{l=1}^{k!} h_{\sigma_l}(\theta)} \ , \qquad i = 1, \ldots, k! \ .$$

If h_{σ_i} is negligible for q, the value η_{σ_i} is almost zero and on the opposite $\eta_{\sigma_i} \approx 1$ indicates a high contribution of h_{σ_i}. The expected relative contribution for $\theta \sim h_{\sigma_c}(\cdot)$ is

$$\mathbb{E}_{h_{\sigma_c}}[\eta_{\sigma_i}(\theta)] = \int_{S_c} \eta_{\sigma_i}(\theta) h_{\sigma_c}(\theta)\, \mathrm{d}\theta$$

estimated by

$$\hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_i}(\theta)] = \frac{1}{M}\sum_{l=1}^{M} \eta_{\sigma_i}(\theta^{(l)}) \ , \qquad \theta^{(l)} \sim h_{\sigma_c}(\cdot) \ . \tag{13}$$

After a possible permutation of the indices, we assume that $\hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}] \geq \cdots \geq \hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}]$, namely that the corresponding $h_{\sigma_1}, \cdots, h_{\sigma_{k!}}$ are in decreasing order of expected contributions. An approximation of q using only the $1 \leq n \leq k!$ most contributing h's is then

$$\tilde{q}_n(\theta) = \frac{1}{k!} \sum_{i=1}^{n} h_{\sigma_i}(\theta) \ , \tag{14}$$

and the mean absolute difference from $q(\theta)$ is approximated by

$$\hat{\phi}_n = \frac{1}{M} \sum_{l=1}^{M} \left| \tilde{q}_n(\theta^{(l)}) - q(\theta^{(l)}) \right| \ . \tag{15}$$

As n closes to $k!$, $\tilde{q}_n$ approaches q and $\hat{\phi}_n$ is almost zero. If this mean absolute difference is below a threshold, τ, $\tilde{q}_n$ is defined as an appropriate approximation for q. The approximate set $\mathfrak{A}(k)$ is then made of $[\sigma_1, \cdots, \sigma_n]$ for the smallest n that satisfies the condition $\hat{\phi}_n < \tau$. Using this derivation, the computation efficiency strictly improves by avoiding negligible h-function computations in q. The thresholds are chosen as a compromise between the approximation quality (n close to $k!$) and computation gain (n far from $k!$).

Note that the above approximation, $\mathfrak{A}(k)$, is determined under the assumption $\theta \sim h_{\sigma_c}(\cdot)$ and that the approximation quality is obviously not guaranteed for θ's not generated from h_{σ_c}. The approximate set $\mathfrak{A}(k)$ is clearly determined by the choice of h_{σ_c} even though the expected size of $\mathfrak{A}(k)$ is fixed due to the perfect symmetry of q over the $k!$ permutations. This means the expected gain in computation is constant regardless to the choice for h_c (and the proof is given in the Appendix). The marginal likelihood estimate using an approximation is given in the next page.

In our double importance sampling scheme, the total number of h-computations required to compute q for all T particles is $(Mk!)+|\mathfrak{A}(k)|(T-M)$ using an approximation. The computational workload reduction rate $\Delta(\mathfrak{A}(k))$ is

$$\Delta(\mathfrak{A}(k)) = \frac{(Mk!) + |\mathfrak{A}(k)|(T-M)}{Tk!} = \frac{M}{T}\left(1 - \frac{|\mathfrak{A}(k)|}{k!}\right) + \frac{|\mathfrak{A}(k)|}{k!} \ . \tag{16}$$

As $|\mathfrak{A}(k)| \to k!$, $\Delta(\mathfrak{A}) \to 1$ and a smaller value indicates a greater reduction in computation. For a given $\mathfrak{A}(k)$, $\Delta(\mathfrak{A})$ converges to $|\mathfrak{A}(k)|/k!$ as $M/T \to 0$ and, in practice, M is much smaller than T.

Algorithm 1: Double importance sampling algorithm

1 Randomly generate a Gibbs sample of size J and construct q as in (11). Label switching in $\{\varphi^{(j)}\}_{j=1}^{J}$ is removed by reference to a MAP approximation, following Jasra et al. (2005).

2 Generate particles from an arbitrary h_{σ_c}, $1 \leq c \leq k!$

$$\theta^{(1)}, \cdots, \theta^{(T)} \sim h_{\sigma_c}(\cdot) \ ,$$

3 Construction of an approximation, $\tilde{q}$:

- **3.1** Subsample M particles from $\{\theta^{(t)}\}_{t=1}^{T}$ and denote them by $\vartheta^{(1)}, \cdots, \vartheta^{(M)}$.
- **3.2** For $l = 1, \ldots, M$, compute $h_{\sigma_1}(\vartheta^{(l)}), \ldots, h_{\sigma_{k!}}(\vartheta^{(l)})$ and $\eta_{\sigma_1}(\vartheta^{(l)}), \ldots, \eta_{\sigma_{k!}}(\vartheta^{(l)})$.
- **3.3** Compute $\hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}], \cdots, \hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}]$ in (13).
- **3.4** Reorder the permutations such that $\hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}(\theta)] \geq \cdots \geq \hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}(\theta)]$.
- **3.5** Initially set $n = 1$ and compute $\tilde{q}_n(\vartheta^{(1)}), \cdots, \tilde{q}_n(\vartheta^{(M)})$ in (14) and ϕ_n in (15). Increase $n = n + 1$ and update $\tilde{q}_n$ and $\hat{\phi}_n$ until $\hat{\phi}_n < \tau$.
- **3.6** Construct a new approximation, $\tilde{q}$, using $h_{\sigma_1}, \ldots, h_{\sigma_n}$.

4 Compute $\tilde{q}(\theta^{(1)}), \ldots, \tilde{q}(\theta^{(T)})$.

5 Substituting $\tilde{q}$ to q in (5), compute $\hat{\mathfrak{E}}$ in (6).

5 Simulation study

Two simulated mixture datasets and real datasets are used to examine the performance of six marginal likelihood estimators. The simulated mixtures are;

- $D_1 : x_1, \ldots, x_{60} \sim 0.3N(-1, 1) + 0.7N(5, 2^2)$
- $D_2 : x_1, \ldots, x_{80} \sim 0.15N(-5, 1) + 0.65N(1, 2^2) + 0.2N(6, 1)$

Both real datasets, called galaxy and fishery datasets, respectively, and provided in Figure 1, have been frequently used in mixture inference as benchmarks (see, e.g. Chib 1995; Frühwirth-Schnatter 2006; Jasra et al. 2005; Richardson and Green 1997; Stephens 2000).

Gaussian and Dirichlet priors are used for the mean μ and proportions λ.

$$\mu \sim N(0, 10^2) \ , \ (\lambda_1, \ldots, \lambda_k) \sim \mathrm{Dir}(1, \ldots, 1) \ .$$

For the variance parameter σ^2, inverse Gamma distributions with two sets of hyper-parameters, $IG(2,3)$ and $IG(2,15)$, are considered. With the second calibration, label switching tends to naturally occur within a Gibbs sequence. The first 5,000 iterations of Gibbs sequences with the length of 10^4 are removed as burn-ins.

First the relative h-function contributions for q with respect to M is numerically tested and the threshold τ is chosen. Marginal likelihoods for mixture models with up to six components are estimated and the performances are compared to standard estimators for all four datasets.

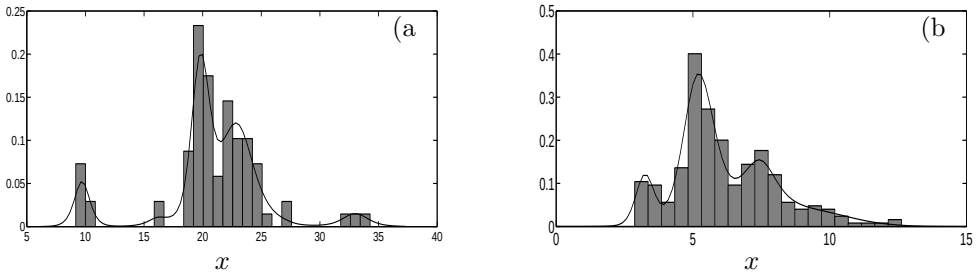

Figure 1: Histogram of the data against estimated six and four Gaussian mixture densities (solid line) for (a) the Galaxy dataset and (b) the Fishery dataset, respectively.

5.1 Determining the approximate set, $\mathfrak{A}(k)$

The choice of τ relates to the quality of the approximation of q and to its computational speed gain. If τ is too small, $\mathfrak{A}(k) \approx \mathfrak{S}(k)$ and there will be hardly any speed gain. A contrario, a very large value of τ will produce a very poor approximation.

The summary of expected relative contributions for D_1 and D_2 is given in Tables 1 and 2. Due to the rounding error, the summation of contribution ratios is not equal to one and the significant contribution functions are easily identified. In MatLab, 10^{-324} is rounded down to 0 and conservatively $\tau = 10^{-324}$ is chosen for all the analyses made in this paper. In Tables 1 and 2, estimates for an approximate set size $|\mathfrak{A}(k)|$ and a error $\hat{\phi}$ are relatively stable against M for a given $J = 100$. When no label switching occurs naturally, q seems to be well approximated using only h_{σ_1} as seen in Table 1. Even a label switching occurs in a Gibbs sequence for a three Normal mixture model, only two functions, h_{σ_1} and h_{σ_2}, contribute for q in Table 2. For the further analysis in this paper, $J = 10^2$, $M = 10^4$ and $\tau = 10^{-324}$ are chosen.

J	M	$\hat{\mathbb{E}}_{h_{\sigma_1}}[\eta]$	$\lvert\mathfrak{A}(k)\rvert$	$\hat{\phi}$
10^2	10^3	$[1, 2\times10^{-83}]$	1	0
10^2	10^4	$[1, 2\times10^{-69}]$	1	0
10^2	10^5	$[1, 5\times10^{-63}]$	1	0

Table 1: Estimates for $\{\hat{\mathbb{E}}_{h_{\sigma_1}}[\eta_i]\}_{i=1}^{k!}$, $\lvert\mathfrak{A}(k)\rvert$ and $\hat{\phi}$ with respect to M for D_1. The prior for a variance parameter is a $IG(2,3)$ distribution.

J	M	$\hat{\mathbb{E}}_{h_{\sigma_1}}[\eta]$	$\lvert\mathfrak{A}(k)\rvert$	$\hat{\phi}$
10^2	10^3	$[1, 8\times10^{-11}, 2\times10^{-59}, 3\times10^{-63}, 7\times10^{-65}, 3\times10^{-90}]$	2	0
10^2	10^4	$[0.9998, 2\times1^{-4}, 3\times10^{-21}, 5\times10^{-50}, 2\times10^{-54}, 4\times10^{-116}]$	2	0
10^2	10^5	$[0.9995, 5\times2^{-4}, 1\times10^{-80}, 9\times10^{-83}, 2\times10^{-94}, 1\times10^{-108}]$	2	0

Table 2: Estimates for $\{\hat{\mathbb{E}}_{h_{\sigma_1}}[\eta_i]\}_{i=1}^{k!}$, $\lvert\mathfrak{A}(k)\rvert$ and $\hat{\phi}$ with respect to M for D_2. The prior for a variance parameter is a $IG(2,15)$ distribution.

5.2 Simulation results

The following five marginal likelihood estimates are compared;

$\widehat{\mathfrak{E}}^{*}_{Ch}$ **:** Chib's method using $T = 5000$ samples with a permutation correction by multiplying $k!$

$\widehat{\mathfrak{E}}_{Ch}$ **:** Chib's method (1), using $T = 5000$ samples which are randomly permuted.

$\widehat{\mathfrak{E}}_{IS}$ **:** Importance sampling estimate (7), using the maximum likelihood estimate (MLE) for $z_1^o, \ldots, z_n^o$ and $T = 10^5$ particles.

$\widehat{\mathfrak{E}}_{DS}$ **:** Dual importance sampling using q in (8), $T = 10^5$ particles and $J = 100$ samples for $q(\theta)$

$\widehat{\mathfrak{E}}^{A}_{DS}$ **:** Dual importance sampling using an approximate in (14), $T = 10^5$ particles, $J = 100$ and $M = 10^4$.

$\widehat{\mathfrak{E}}_{BS}$ **:** Bridge sampling (3), using $L = M = 6000$ samples and $J = 4000$ samples for $q(\theta)$ in (4) via 10 iterations. Here, label switching is imposed in hyperparameters $\{\theta^{(j)}, z^{(j)}\}_{j=1}^{J}$.

The marginal likelihood estimates in log scales ($\log(\widehat{\mathfrak{E}})$) and the effective sample size (ESS) ratios ($r = \mathrm{ESS}/T$) are summarised based on 50 replicates.

Simulated mixture dataset

Mixture models of two and three components are fitted to D_1 and D_2 respectively. With a prior $IG(2,3)$ on the variance parameters, label switching does not occur and all estimates relatively coincide, except for $\log(\widehat{\mathfrak{E}}_{IS})$ in Figure 2. When label switching naturally occurs in the Gibbs sequence under the prior $IG(2,15)$, the $\log(\widehat{\mathfrak{E}}^*_{Ch})$ values also slightly disagree with the other estimates in Figure 3. This unsurprisingly indicates an incorrect correction for label switching through a multiplication by $k!$ as Neal (1999); Frühwirth-Schnatter (2006) and Marin and Robert (2008) reported.

The smallest effective sample size is observed for $r(\widehat{\mathfrak{E}}_{IS})$ and this relates to a larger variation of $\log(\widehat{\mathfrak{E}}_{IS})$ values when compared with estimates using dual importance sampling. The estimate $\log(\hat{\mathfrak{E}}_{IS})$ is an integral approximation of $f(x|\theta)\pi(\theta)$ over the support of $\pi(\cdot|x,z^o)$ in which z^o maximises the likelihood. As the number of components increases, this supporting domain may not be large enough to fit $f(x|\theta)\pi(\theta)$ and poor estimates likely result. For $k=3$, poor marginal likelihood estimates induced by this poor importance function clearly show in both Figures.

In general, dual importance sampling, Chib's method with a random permutation, and bridge sampling appear as the most suitable estimators for this mixture model simulation study. In particular, for dual importance sampling, no significant difference in approximations of $\log(\mathfrak{E})$ and in effective sample sizes is observed, when using a suitable approximate for q. Moreover, the mean sizes of $\mathfrak{A}(k)$ in Table 3 show that $\log(\mathfrak{E})$ can be estimated with a lesser computational workload, with a reduction rate of 0.33.

D	k	$k!$	$\lvert\overline{\mathfrak{A}_1(k)}\rvert$	$\overline{\Delta}(\mathfrak{A}_1)$	$\lvert\overline{\mathfrak{A}_2(k)}\rvert$	$\overline{\Delta}(\mathfrak{A}_2)$
D_1	2	2	1	0.5005	2	1.0000
D_2	3	6	2	0.3340	2.14	0.3573

Table 3: Mean estimates for the approximation set size, $|\mathfrak{A}(k)|$, and the computation reduction factor, $\Delta(\mathfrak{A})$, in (16) for D_1 and D_2. The subscripts 1 and 2 indicate the results using the priors $\sigma^2 \sim IG(2,3)$ and $\sigma^2 \sim IG(2,15)$, respectively.

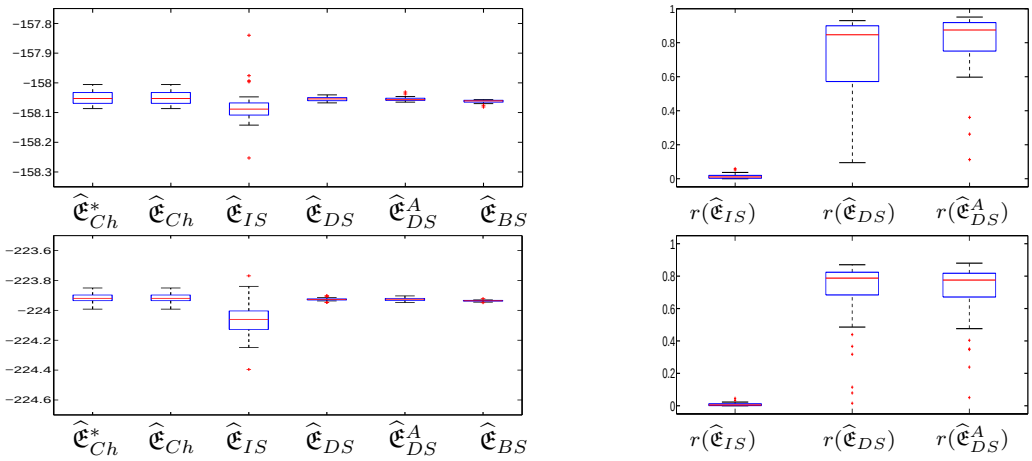

Figure 2: Boxplots of marginal likelihood estimates in log scale *(left)* and effective sample sizes ratios *(right)*. Note that for conciseness' sake, $\log(\widehat{\mathfrak{E}})$ is replaced with $\widehat{\mathfrak{E}}$ in the label captions. Mixture models with *(top)* two and *(bottom)* three Gaussian components are fitted to D_1 and D_2, respectively. The prior on σ^2 is the distribution $IG(2, 3)$.

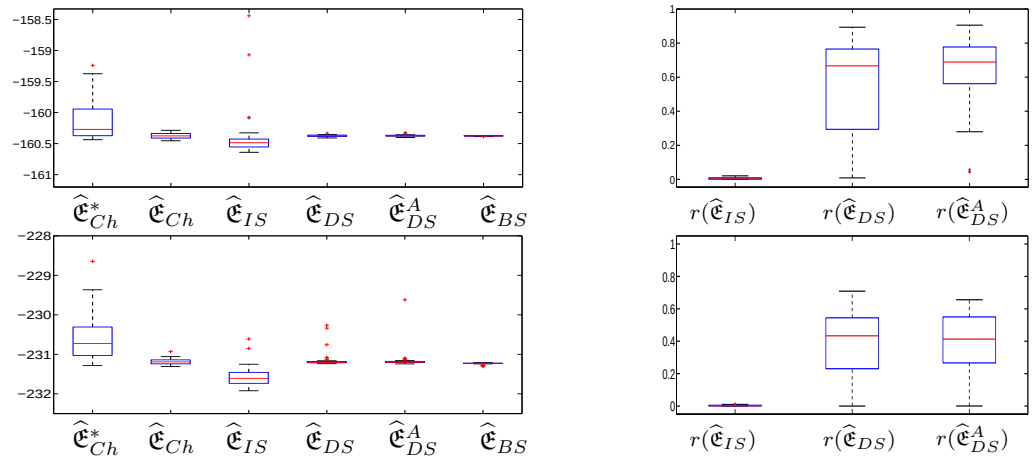

Figure 3: Boxplots of marginal likelihood estimates in log scale *(left)* and effective sample sizes ratios *(right)*. Note that for conciseness' sake, $\log(\widehat{\mathfrak{E}})$ is replaced with $\widehat{\mathfrak{E}}$ in the label captions. Mixture models with *(top)* two and *(bottom)* three Gaussian components are fitted to D_1 and D_2, respectively. The prior on σ^2 is the distribution $IG(2, 15)$.

Galaxy and fishery dataset

The priors suggested by Richardson and Green (1997) are used for our simulation study:

$$\begin{aligned} \mu &\sim N(\bar{x}, R^2/4) \\ \sigma_i^2 &\sim IG(2, \beta)\ ,\ i = 1, \cdots, k \\ \beta &\sim G(0.2, 10/R^2) \\ \lambda_1, \ldots, \lambda_k &\sim \text{Dirichlet}(1, \cdots, 1) \end{aligned}$$

Here $\bar{x}$ and R are the median and the range of x, respectively.

Normal mixture models are fitted to both datasets and estimates for $\log(\mathfrak{E}(k))$ are summarised in Figures 4 and 5. In general, a similar phenomenon in the behaviour of $\log(\hat{\mathfrak{E}}(k))$ and $r(\hat{\mathfrak{E}})$ in terms of the methods is observed. Unless the components are clearly separated, $|\overline{\mathfrak{A}(k)}| \approx 1$, $\log(\hat{\mathfrak{E}}^*_{Ch})$ estimate is biased due to a wrong permutation correction. The importance sampling approach with a poor q results an integration approximation over a limited support instead of the full support of $f(x|\theta)\pi(\theta)$ and the estimate gets poorer as k increases. Even in the case of an eight dimensional posterior (when $k = 3$), this estimator already significantly suffers.

As k increases, the number of effective sample size is reduced exponentially and the variation in $\log(\hat{\mathfrak{E}})$ estimates increases. Particularly, as k increases up to six, a variation increase in $\log(\hat{\mathfrak{E}}_{Ch})$ values with k is significantly greater than $\log(\hat{\mathfrak{E}}_{DS})$ and $\log(\hat{\mathfrak{E}}^A_{DS})$ values.

The reduction in computation due to the use of an approximation is observed for all simulations and the reduction rate varies by cases as shown in Table 4. Particularly, when $k = 4$ and $k = 6$, normal components for the galaxy data tend to have long flat tails and have higher chances to overlap each other. Consequently, the computation workload reduction is less than a model using a smaller number of components.

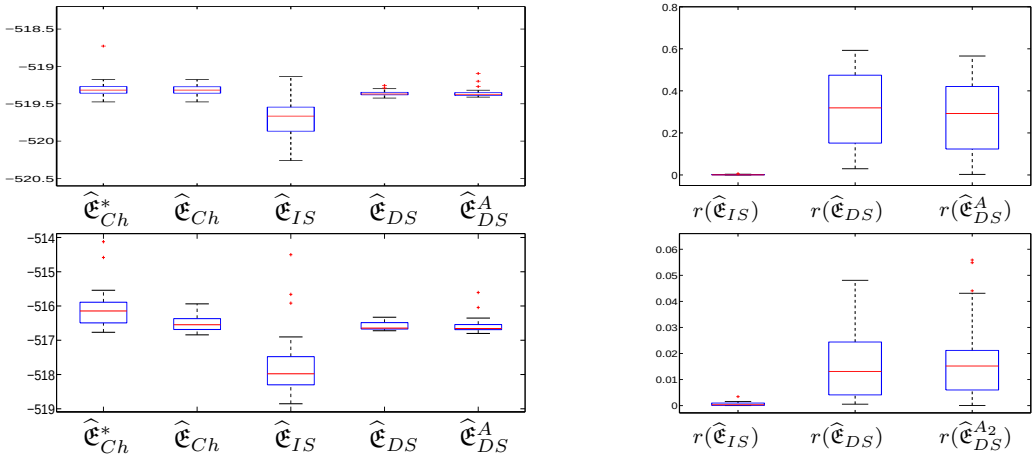

Figure 4: Boxplots of marginal likelihood estimates in log scale *(left)* and effective sample sizes ratios *(right)*. Note that for conciseness' sake, $\log(\widehat{\mathfrak{E}})$ is replaced with $\widehat{\mathfrak{E}}$ in the label captions. Mixture models with *(top)* three and *(bottom)* four Gaussian components are fitted to the fishery dataset.

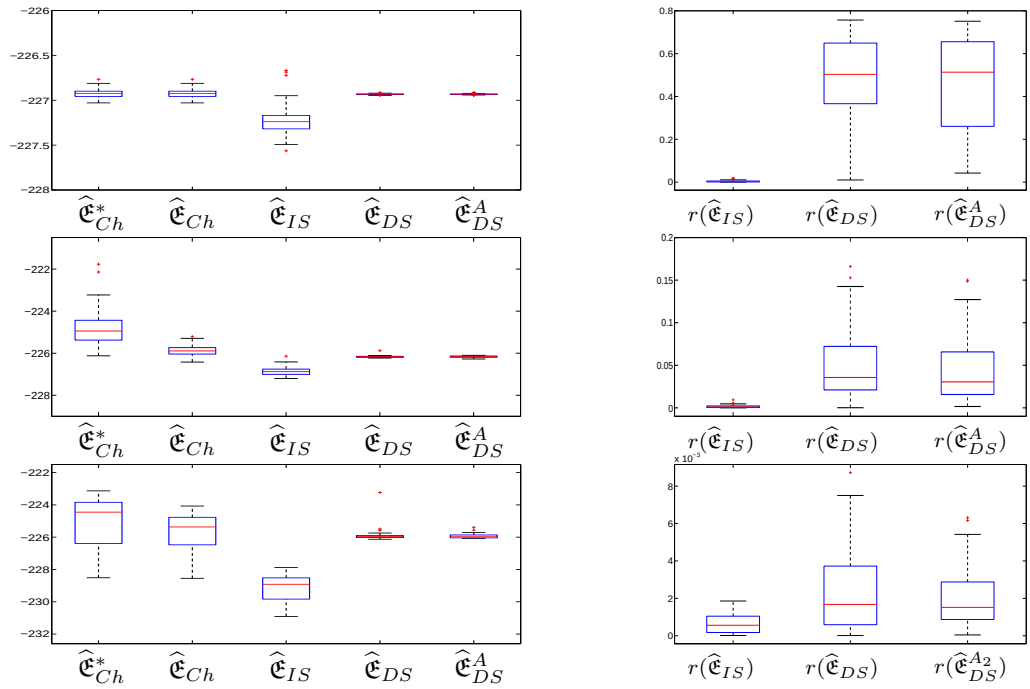

Figure 5: Boxplots of marginal likelihood estimates in log scale *(left)* and effective sample sizes ratios *(right)*. Note that for conciseness' sake, $\log(\widehat{\mathfrak{E}})$ is replaced with $\widehat{\mathfrak{E}}$ in the label captions. Mixture models with *(top)* three and *(bottom)* four Gaussian components are fitted to the galaxy dataset.

k	k!	$\overline{\lvert\mathfrak{A}(k)\rvert}$	$\overline{\Delta}(\mathfrak{A})$
3	6	1.0000	0.1675
4	24	2.7333	0.1148

(a) Fishery data

k	$k!$	$\overline{\lvert\mathfrak{A}(k)\rvert}$	$\overline{\Delta}(\mathfrak{A})$
3	6	1.000	0.1675
4	24	16.8333	0.7188
6	720	298.1200	0.4146

(b) Galaxy data

Table 4: Mean estimates of approximate set sizes, $|\mathfrak{A}(k)|$, and the computation reduction ratio, $\Delta(\mathfrak{A})$ in (16) for (a) fishery and (b) galaxy datasets.

5.3 Importance function choice

In importance sampling, the importance function q highly influences to estimates. For mixture models, key aspects for the choice of q are (a) finding a support of q that includes the support of the posterior density and (b) its capability of inducing label switching. Two importance functions, (4) and (8), allow for label switching and the corresponding estimates are (9) and (10) respectively. Theoretically we proved that the asymptotic variance for (10) is smaller than (9) and this is numerically examined in this

section.

Mixture models using three and four Gaussian components are fitted for the simulated data D_2 and galaxy data, respectively. The estimates in log-scale and the effective sample size rates are summarised in Figure 6. The estimate (9) is denoted by $\hat{\mathfrak{E}}_J$ and J is the number of conditional functions used for q. When $J = 100$, the variations of $\hat{\mathfrak{E}}_J$ values are greater than $\hat{\mathfrak{E}}_{DS}$ and the effective samples sizes are smaller. For the galaxy data, the variance and the effective samples sizes of those estimates become similar when $J = 3000$ for the estimator (9). These observations obviously support our theoretical result.

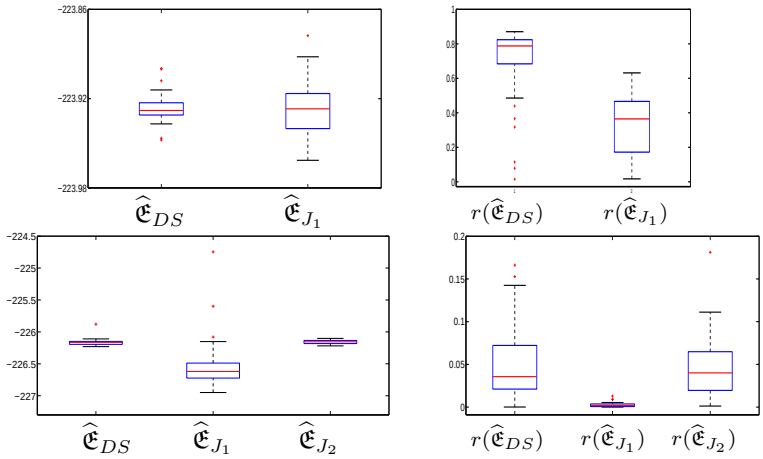

Figure 6: Boxplots of marginal likelihood estimates in log scale *(left)* and effective sample sizes ratios *(right)*. Note that for conciseness' sake, $\log(\widehat{\mathfrak{E}})$ is replaced with $\widehat{\mathfrak{E}}$ in the label captions. Mixtures of three Gaussians *(top)* and four Gaussian components *(bottom)* are fitted to D_2 and the galaxy dataset, respectively. The importance sampling approximation (4) using J Gibbs samples for q is denoted by $\hat{\mathfrak{E}}_J$ and $J_1 = 100$ and $J_2 = 3000$.

6 Discussion

This paper considers the evidence approximations using the importance sampling for mixture models and some of known challenges due to a complexity of multimodal posterior density arising with k. For importance sampling based estimators, it is essential that a support of an importance function is large enough compared to the support of the posterior density. Particularly for mixture models, an inappropriate permutation approximation for an importance function is likely to result an unsuitable support hence, a poor estimate.

For our investigation the common priors for all mixture components are assumed and, consequently the posterior and marginal posterior densities are made up with $k!$ symmetrical terms. Two marginal likelihood estimators are proposed and tested along other existing estimators. The first approach uses the permutation representations of $\pi(\cdot|x, z^o)$ with a pointwise MLE, z^o, for an importance function. Obviously due to a poor support of an importance function, this approach performs poorly in our simulation studies. Another poor estimates using the Chib's method are observed when the permutations are incorrectly approximated.

Secondly the importance function is constructed adapting the Rao-Blackwellised twice to approximate a multimodal marginal posterior density and, this is called the dual importance sampling. Theoretically and practically it is demonstrated that the dual importance sampling is a compatible estimator particularly for mixture model and even a higher efficiency gained compared to an importance function by Frühwirth-Schnatter (2001). Moreover, using a suitable approximation for an importance function, its exponential increasing computational workload with k can be reduced. The underlying idea is to avoid negligible function computations using the perfect symmetry $k!$ terms of posterior density. As posterior modes are well separated, a gain by the use of an approximation is greater and more are overlapped, a less gain is resulted.

Borrowing a similar approach in Chib (1996), the dual importance sampling can be extended to the cases in which conditional densities in a Gibbs sampling representation are not in closed forms. However it is suffered from the curse of dimensionality just like any other estimators based on importance sampling and, this is observed in our simulation studies.

Alternative model evidence approximations are well summarised Friel and Wyse (2012). Among them, the ensemble Monte Carlo approach is that samples from local ensembles are extensions or compositions of the original. For example, parallel tempering Monte Carlo method, ensembles of states (Neal), among others. Extending this idea, Bayes factor approximations were proposed such as annealed importance sampling by Neal (2001) and power posteriors by Friel and Pettitt (2008). Further investigation is needed to characterise their performances for mixture models with respect to the label

switching and k.

Appendix

Let $\sigma_c(\theta)$ be a particle θ in a particular permutation representation, σ_c. Recalling the equation (12) the followings are true for any σ_m,

$$\begin{aligned} h_{\sigma_i}(\sigma_c(\theta)) &= \frac{1}{J}\sum_{j=1}^{J}\pi(\sigma_c(\theta)|\sigma_i(\varphi^{(j)}),x) \\ &= \frac{1}{J}\sum_{j=1}^{J}\pi(\sigma_m\sigma_c(\theta)|\sigma_m\sigma_i(\varphi^{(j)}),x) = h_{\sigma_m\sigma_i}(\sigma_m\sigma_c(\theta)) \end{aligned}$$

and

$$\eta_{\sigma_i}(\sigma_c(\theta)) = \frac{h_{\sigma_i}(\sigma_c(\theta))}{k!q(\theta)} = \frac{h_{\sigma_m\sigma_i}(\sigma_m\sigma_c(\theta))}{k!q(\theta)} = \eta_{\sigma_m\sigma_i}(\sigma_m\sigma_c(\theta)) \ .$$

Suppose that $h_{\sigma_1},\cdots,h_{\sigma_{k!}}$ are in the order of $\hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}] \geq \cdots \geq \hat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}]$. Then, the order of $h_{\sigma_m\sigma_1},\cdots,h_{\sigma_m\sigma_{k!}}$ corresponds to $\hat{\mathbb{E}}_{h_{\sigma_m\sigma_c}}[\eta_{\sigma_m\sigma_1}] \geq \cdots \geq \hat{\mathbb{E}}_{h_{\sigma_m\sigma_c}}[\eta_{\sigma_m\sigma_{k!}}]$ and h's are in a decreasing contribution order for q assuming that particles are generated from $h_{\sigma_m\sigma_c}$. If $\sigma_m\sigma_c = \sigma_c$, the order of h's for both cases are identical. Otherwise, they are not identical however, the expected absolute errors using the n most contributing h-functions are equal,

$$\tilde{q_n}(\sigma_c(\theta)) = \frac{1}{k!}\sum_{i=1}^{n} h_{\sigma_i}(\sigma_c(\theta)) = \frac{1}{k!}\sum_{i=1}^{n} h_{\sigma_m\sigma_i}(\sigma_m\sigma_c(\theta)) = \tilde{q_n}(\sigma_m\sigma_c(\theta))$$

and

$$\mathbb{E}_{h_{\sigma_c}}(\phi_n) = \int \left|\frac{1}{k!}\sum_{i=1}^{n} h_{\sigma_i}(\theta) - q(\theta)\right| h_{\sigma_c}(\theta)d\theta = \int \left|\frac{1}{k!}\sum_{i=1}^{n} h_{\sigma_m\sigma_i}(\theta) - q(\theta)\right| h_{\sigma_m\sigma_c}(\theta)d\theta = \mathbb{E}_{h_{\sigma_m\sigma_c}}(\phi_n) \ .$$

The approximate set $\mathfrak{A}(k)$ is chosen by

$$\min_n\{\mathbb{E}_{h_{\sigma_c}}[\phi_n] < \tau\} = \min_n\{\mathbb{E}_{h_{\sigma_m\sigma_c}}[\phi_n] < \tau\} \ .$$

Any permutation representation for θ is constructed using $\sigma_m\sigma_c$ by the choice for σ_m hence, the expected value for $|\mathfrak{A}(k)|$ is same for any h_c in Algorithm 1.

References

Berkhof, J., van Mechelen, I., and Gelman, A. (2003). "A Bayesian approach to the selection and testing of mixture models." *Statistical Sinica*, 13(3): 423–442.

Celeux, G., Hurn, M., and Robert, C. (2000). "Computational and inferential difficulties with mixture posterior distributions." *Journal of American Statistical Association*, 95(3): 957–979.

Chen, M.-H., Shao, Q., and Ibrahim, J. (2000). *Monte Carlo methods in Bayesian computation*. Springer Series in Statistics, 1 edition.

Chib, S. (1995). "Marginal likelihoods from the Gibbs output." *Journal of the American Statistical Association*, 90: 1313–1321.

— (1996). "Calculating posterior distributions and modal estimates in Markov mixture models." *Journal of Econometrics*, 75: 79–97.

Chopin, N. (2002). "A sequential particle filter method for static models." *Biometrika*, 89(3): 539–552.

Chopin, N. and Robert, C. (2010). "Properties of Nested sampling." *Biometrika*, 97: 741–755.

Dempster, A., Laird, N., and Rubin, D. (1997). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of Royal Statistical Society, Series B*, 39(1): 1–38.

Diebolt, J. and Robert, C. (1994). "Estimation of finite mixture distributions through Bayesian sampling." *Journal of Royal Statistical Society, Series B*, 56: 363–375.

Friel, N. and Pettitt, A. (2008). "Marginal likelihood estimation via power posteriors." *Journal of the Royal Statistical Society, Series B*, 70: 589607.

Friel, N. and Wyse, J. (2012). "Estimating the evidence a review." *Statistica Neerlandica*, 66(3): 288–308.

Frühwirth-Schnatter, S. (2001). "Markov Chain Monte Carlo estimation fo classical and dynamic switching and mixture models." *Journal of the American Statistical Association*, 96: 194–209.

— (2004). "Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques." *Econometrics*, 7: 143–167.

— (2006). *Finite mixture and Markov switching models*. Springer-Verlag, New York.

Gelfand, A. and Smith, A. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85: 398–409.

Gelman, A. and Meng, X. (1998). "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling." *Statistical Science*, 13: 163–185.

Green, P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 85(4): 711–732.

Jasra, A., Holmes, C., and Stephens, D. (2005). "Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling." *Statistical Science*, 20(1): 50–67.

Jeffreys, H. (1939). *Theory of Probability*. Oxford, The Clarendon Press, 1 edition.

Marin, J.-M. and Robert, C. (2008). "Approximating the marginal likelihood in mixture models." *Bulletin of the Indian Chapter of ISBA*, 1: 2–7.

Meng, X. and Schilling, S. (2002). "Warp Bridge sampling." *American Statistical Association*, 11(3): 552–586.

Meng, X. and Wong, W. (1996). "Simulating ratios of normalizing constants via a simple identity." *Statistica Sinica*, 6: 831–860.

Mira, A. and Nicholls, G. (2004). "Bridge estimation of the probability density at a point." *Statistica Sinica*, 14: 603–612.

Neal, R. (1999). "Erroneous results in "Marginal likelihood from the Gibbs output"." Http://www.cs.toronto.edu/ radford/chib-letter.html.

— (2001). "Annealed importance sampling." *Statistics and Computing*, 11: 125139.

Newton, M. and Raftery, A. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap." *Journal of Royal Statistical Society, Series B*, 96(1): 3–48.

Richardson, S. and Green, P. (1997). "On Bayesian analysis of mixtures and with an unknown number of components." *Journal of the Royal Statistical Society, Series B*, 59(4): 731–792.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2 edition.

Servidea, J. (2002). "Bridge sampling with dependent random draws:techniques and strategy." Ph.D. thesis, Department of Statistics, The University of Chicago.

Skilling, J. (2007). "Nested sampling for Bayesian computations." *Bayesian Analysis*, 1(4): 833–859.

Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of Royal Statistical Society, Series B*, 62: 795–809.

Voter, A. (1985). "A Monte Carlo method for determining free-energy differences and transition state theory rate constants." *Journal of Chemical Physics*, 82: 1890–1899.

Acknowledgments

The author(s) wish to thank CREST, INSEE, Paris, for its hospitality during this collaboration, as well as CEREMADE, UMR 7534, CNRS and Universit Paris-Dauphin, and Auckland University of Technology for their financial support of the first author's visits in Paris. The second author's research is supported by a 2010–2015 senior Institut Universitaire de France grant, which he gratefully acknowledges.