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 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. 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

Consider a sample $\mathbf{x} = \{x_1, \cdots, x_{n_x}\}$ that is a realisation of a random sample (univariate or multivariate) from a finite mixture of k distributions

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

where the component weights $\boldsymbol{\lambda} = \{\lambda_i\}_{i=1}^{k}$ are non-negative and sum to 1. The collection of the component-specific parameters is denoted $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^{k}$ and the collection of all parameters by $\theta = \{\boldsymbol{\lambda}, \boldsymbol{\xi}\}$. As is now standard (Marin et al. 2005) each observation x_j from the sample can be assumed to originate from a specific if unobserved component of f_k, denoted z_i, and the mixture inference problem can then be analysed as a missing

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

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

data model, with discrete missing data $\mathbf{z} = \{z_1, \ldots, z_{n_x}\}$, such that

$$x_j|\mathbf{z} \sim f(x_j|\xi_{z_j})\,, \qquad \text{independently for } j = 1, \cdots, n_x\,.$$

The conditional distribution of $z_j \in [1, \ldots, k]$ is then given by

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

This interpretation of the mixture model leads to a natural clustering of the observations according to their label and the cluster associated with the mixture component i provides information about λ_i and ξ. In particular, when the full conditional distribution of the parameter θ is available in closed form, conditional simulation from $\pi(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathbf{x}, \mathbf{z})$ becomes straightforward (see Diebolt and Robert (1994)).

In a Bayesian mixture modelling setup, the goal is to perform inference on the parameter θ and the posterior distribution $\pi(\theta|\mathbf{x})$ is usually approximated via MCMC methods. The likelihood function $p_k(\mathbf{x}|\theta)$ is both available and invariant under permutations of the component indices. If an exchangeable prior is chosen on $(\boldsymbol{\lambda}, \boldsymbol{\xi})$, the posterior density reproduces the likelihood invariance and component labels are not identifiable. This phenomenon is called *label switching* and is well-studied in the literature (Celeux et al. 2000; Stephens 2000b; Jasra et al. 2005). From a simulation perspective, label switching induces multimodality in the target and while it is desirable that a simulated Markov chain targeting the posterior explores all of the $k!$ symmetric modes of the posterior distribution, most samplers fail to switch between modes (Celeux et al. 2000). For instance, when using a data augmentation scheme, which is a form of Gibbs sampler adapted to missing data problems (Robert and Casella 2004), the Markov chain very slowly if ever switches between the symmetric modes. Therefore, since the chain only explores a certain region of the support of the multimodal posterior, estimates based on the simulation output are necessarily biased. When label switching is missing from the MCMC output, it can be simulated by modifying the MCMC sample (see Frühwirth-Schnatter (2001); Papastamoulis and Roberts (2008); Papastamoulis and Iliopoulos (2010)).

A different perspective on the label switching phenomenon is inferential. Indeed, point estimates of the component-wise parameters are harder to produce when the Markov chain moves freely between modes. To achieve component-specific inference and give a meaning to each component, relabelling methods have been proposed in the literature (see Richardson and Green (1997); Celeux et al. (2000); Stephens (2000b); Jasra et al. (2005); Marin and Robert (2007); Geweke (2012); Rodriguez and Walker (2014) and others). An R-package, `label.switching` (Papastamoulis 2013), incorporates some of those label switching removing methods.

Evaluating the number of components k is a special case of model comparison, which can be conducted by comparing the *posterior probabilities of the models*. Those probabilities are in turn computed via the marginal likelihoods $\mathfrak{E}(k)$, also known as model evidences (Richardson and Green 1997)

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

where $\pi_k(\theta)$ is the selected prior for the k-component mixture. (We assume here that it is exchangeable wrt its components.) Recall that the ratio of evidences is a Bayes factor and is properly scaled to be readily compared to 1 (Jeffreys 1939). When a large collection of values of k is considered for model comparison, sophisticated MCMC methods have been developed to bypass computing evidences (Richardson and Green 1997; Stephens 2000a), even though those are estimated as a byproduct of the methods. Alternatively, estimating the number of components can also proceed from a Bayesian nonparametric (BNP) modelling, which assumes an infinite number of components and then evaluates the non-empty components implicitly through partitioning data, using for instance the Chinese restaurant process (Antoniak 1974; Escobar and West 1995; Rasmussen 2000). This however requires a modification of the prior modelling and we will not cover it in this paper, which reassesses Monte Carlo ways of approximating the evidence.

The difficulty with approaches using $\mathfrak{E}(k)$ is that the quantity often cannot directly and reliably be derived from simulations from the posterior distribution $\pi(\theta|\mathbf{x})$ (Newton and Raftery 1994). The quantity has been approximated using dedicated methods such harmonic means (Satagopan et al. 2000; Raftery et al. 2006), importance sampling (Rubin 1987, 1988; Gelman and Meng 1998), bridge sampling (Meng and Wong 1996; Meng and Schilling 2002), Laplace approximation (Tierney and Kadane 1986; DiCiccio et al. 1997), stochastic substitution (Gelfand and Smith 1990; Chib 1995, 1996), nested sampling (Chopin and Robert 2010), Savage-Dickey representations (Verdinelli and Wasserman 1995; Marin and Robert 2010b) and erroneous implementations of the Carlin and Chib algorithm (Carlin and Chib 1995; Scott 2002; Congdon 2006; Robert and Marin 2008). Comparative studies of those methods are found in Marin and Robert (2010a) and Ardia et al. (2012).

In the specific case of mixtures, the invariance of the posterior density under an arbitrary relabelling of the mixture components must be exhibited by simulations and approximations to achieve a valid estimate of $\mathfrak{E}(k)$ as discussed in Neal (1999); Berkhof et al. (2003); Marin and Robert (2008). This often leads to computationally intensive steps in approximation methods, especially when k is large, and it is the purpose of this paper to provide a partial answer to this specific issue.

We consider here two estimators of $\mathfrak{E}(k)$, both based on importance sampling (IS). One is a version of Chib's estimator and the second one a novel representation called *dual importance sampling*. Our importance construction aims to better approximate the posterior distribution both around a specific local mode and at the corresponding $(k! - 1)$ symmetric modes of the posterior distribution. A particular mode is first approximated based on (i) a point estimate and (ii) Rao–Blackwellisation from a Gibbs sequence. Then, the corresponding local density approximate is permuted to reach all modes. We demonstrate here that dual importance sampling is comparable to our benchmark method, Chib's approach. Taking advantage of the symmetry in the posterior distribution, we show how to reduce computational demands by approximating our importance function.

The paper starts with recalling the approximation techniques of Chib's method and bridge sampling in Section 2. In Section 3, importance sampling is considered, including our choices of importance functions. Our importance function approximate approach is introduced in Section 4. Simulation studies using both simulated and benchmark datasets, namely the galaxy and fishery datasets used in Richardson and Green (1997) are reported in Section 5, and the paper concludes with a short discussion in Section 6.

2 Standard evidence estimators

2.1 Chib's estimator and corrections

In this paper, the reference estimator of evidence is Chib's(1995) method and is derived from rewriting Bayes' theorem

$$\widehat{\mathfrak{E}}(k) = m_k(\mathbf{x}) = \frac{\pi_k(\theta^o)p_k(\mathbf{x}|\theta^o)}{\pi_k(\theta^o|\mathbf{x})} \tag{1}$$

where θ^o is any plug-in value for θ. When $\pi_k(\theta^o|\mathbf{x})$ is not available in closed form, the Gibbs sampling decomposition allows a Rao–Blackwellised approximation (Gelfand and Smith 1990; Robert and Casella 2004)

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

where $(\mathbf{z}^t)_{t=1}^T$ is a Markov chain with stationary distribution $\pi_k(\mathbf{z}|\mathbf{x})$. The appeal of this estimator, when available, is that it constitutes a non-parametric density estimator converging at a regular parametric rate.

It is now an accepted fact that label switching is necessary for the above Rao–Blackwellised $\hat{\pi}_k(\theta^o|\mathbf{x})$ to converge. When $(\mathbf{z}^1, \cdots, \mathbf{z}^T)$ only explores part of the modes of the posterior, this estimator is biased, generally missing the target quantity $\log(m_k(\mathbf{x}))$ by a factor of order O(log $k!$), with no simple correction factor (Neal 1999). Berkhof et al. (2003) later suggested a generic correction by averaging $\widehat{\pi}_k(\theta^o|\mathbf{x})$ over all possible permutations of the labels, hence forcing "perfect" label switching. The resulting approximation is expressed as

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

where $\mathfrak{S}_k$ denotes the set of the $k!$ permutations of $\{1, \ldots, k\}$ and σ is one of those permutations. Note that the above correction can also be rewritten as

$$\tilde{\pi}_k(\theta^o|\mathbf{x}) = \frac{1}{Tk!}\sum_{\sigma\in\mathfrak{S}_k}\sum_{t=1}^{T} \pi_k(\sigma(\theta^o)|\mathbf{x}, \mathbf{z}^t)\,, \tag{2}$$

using a notational shortcut $\sigma(\theta^o)$ meaning that the components of θ^o are then switched according to the permutation σ. This representation may induce computational gains since only $k!$ versions of $\sigma(\theta^o)$ need to be stored.

While Chib's representation has often been advocated as a highly stable solution for computing the evidence in mixture models, which is why we selected it as our reference, alternative solutions abound within the literature, including nested sampling (Skilling 2007; Chopin and Robert 2010), reversible jump MCMC (Green 1995; Richardson and Green 1997), and particle filtering (Chopin 2002).

2.2 Bridge Sampling

Meng and Wong (1996) proposed a sample–based method to compute a ratio of normalizing constants of two densities with common support. The method is well-suited to estimate the marginal likelihood (Frühwirth-Schnatter 2001, 2004) and used as a point posterior estimate for Chib's method (Mira and Nicholls 2004). Considering a normalised density q and the unnormalized posterior distribution $\pi_k^*(\theta|\mathbf{x}) = \pi_k(\theta)p_k(\mathbf{x}|\theta)$, the bridge sampling identity is given by

$$\widehat{\mathfrak{E}}(k) = \frac{\mathbb{E}_{q(\theta)}[\alpha(\theta)\pi_k^*(\theta|\mathbf{x})]}{\mathbb{E}_{\pi_k(\theta|\mathbf{x})}[\alpha(\theta)q(\theta)]}\,,$$

for an arbitrary function α (provided all expectations are well-defined, Chen et al. 2000). The (formally) optimal choice for α (Meng and Wong 1996) leads to the following iterative estimator

$$\widehat{\mathfrak{E}}^{(t)}(k) = \widehat{\mathfrak{E}}^{(t-1)}(k)\,\frac{M_1^{-1}\displaystyle\sum_{l=1}^{M_1} {}^{\hat{\pi}_{t-1}(\tilde{\theta}^l|\mathbf{x})}\big/_{M_1 q(\tilde{\theta}^l)+M_2\hat{\pi}_{t-1}(\tilde{\theta}^l|\mathbf{x})}}{M_2^{-1}\displaystyle\sum_{m=1}^{M_2} {}^{q(\hat{\theta}^m)}\big/_{M_1 q(\hat{\theta}^m)+M_2\hat{\pi}_{t-1}(\hat{\theta}^m|\mathbf{x})}} \tag{3}$$

where $\hat{\pi}_{t-1}(\theta|\mathbf{x}) = \pi_k^*(\theta|\mathbf{x})/\widehat{\mathfrak{E}}^{(t-1)}(k)$. Here, $(\tilde{\theta}^1, \ldots, \tilde{\theta}^{M_1})$ and $(\hat{\theta}^1, \ldots, \hat{\theta}^{M_2})$ are samples from $q(\theta)$ and $\pi_k(\theta|\mathbf{x})$ respectively.

The convergence of bridge sampling (with the above optimal α) is trivial when $\pi_k^*(\theta|\mathbf{x})$ and $q(\theta)$ share a sufficiently large portion of their supports. If the support intersection is too small, the resulting bridge sampling estimator may end up with an infinite variance (Voter 1985; Servidea 2002). Improvements of the algorithm, like path sampling (Gelman and Meng 1998), a simple location shift of the proposal distribution (Voter 1985), and a warp bridge sampling (Meng and Schilling 2002), have been proposed.

In the specific case of the mixture posterior distribution, the parameter θ can be split in $\boldsymbol{\lambda}$ and k further blocks $\xi_1, \ldots, \xi_k$ in the Gibbs sampling steps. The output samples from the Gibbs sampler are denoted by $\{\theta^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^{J_1}$, where the $\mathbf{z}^{(j)}$'s are the component allocation vectors associated with the observations $\mathbf{x}$. For bridge sampling, Frühwirth-Schnatter (2004) suggested using a Rao–Blackwellised function $q(\theta) = q(\boldsymbol{\lambda}, \boldsymbol{\xi})$

of the form

$$\begin{aligned} q(\theta) &= \frac{1}{J_1}\sum_{j=1}^{J_1} \pi_k(\theta|\theta^{(j)},\mathbf{z}^{(j)},\mathbf{x}) \qquad (4) \\ &= \frac{1}{J_1}\sum_{j=1}^{J_1} p(\boldsymbol{\lambda}|\mathbf{z}^{(j)})\prod_{i=1}^{k} p(\xi_i|\boldsymbol{\xi}^{(j)},\mathbf{z}^{(j)},\mathbf{x}) \end{aligned}$$

assuming $\{\theta^{(j)},\mathbf{z}^{(j)}\}_{j=1}^{J_1}$ is well-mixed, followed by switching the labels of the simulations from the posterior distribution (Frühwirth-Schnatter 2001). Frühwirth-Schnatter (2004) demonstrated that the iterative bridge sampling estimator (3), using (4) as $q(\cdot)$, converges relatively quickly, in about $t = 10$ iterations, even with different starting values.

3 New importance sampling estimators

If $q(\theta)$ is an importance function with support S_q, generating a sample $\boldsymbol{\theta} = (\theta^{(1)},\ldots,\theta^{(T)})$ from $q(\theta)$ leads to the evidence approximation

$$\widehat{\mathfrak{E}}(k) = \frac{1}{T}\sum_{t=1}^{T}\frac{\pi_k(\theta^{(t)})p_k(\mathbf{x}|\theta^{(t)})}{q(\theta^{(t)})} \overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\omega(\theta^{(t)})\,. \qquad (5)$$

To provide a good approximation, the support of $q(\theta)$ must overlap the support of the posterior distribution, which is both symmetric under permutations and multimodal. In this sense, a Rao–Blackwellised estimate like (4) is a natural solution for the choice of q, despite the drawback that J_1 increases “factorially” fast with k due to the permutations over $\{\theta^{(j)},\mathbf{z}^{(j)}\}_{j=1}^{J_1}$ (Frühwirth-Schnatter 2004; Frühwirth-Schnatter 2006).

In the following sections, the parameter θ is decomposed into $(k+1)$ blocks $\theta = (\boldsymbol{\lambda},\xi_1,\ldots,\xi_k)$. Note that ξ_i is a component-wise block, most often a vector. Two types of importance functions, based on the product of marginal posterior distributions, will be considered. The usefulness and details of the product of block marginal posterior distributions are well summarised in Perrakie et al. (2014).

3.1 A plug-in proposal

Using a very simple Rao–Blackwell argument inspired from Chib’s representation, a natural importance function is

$$q(\theta) = \pi_k(\theta|\mathbf{z}^o,\theta^o,\mathbf{x}).$$

Samples are generated from the posterior distribution conditional on a given completion vector $\mathbf{z}^o$, which is usually taken equal the MAP (maximum a posteriori) or the marginal MAP estimate of $\mathbf{z}$ derived from MCMC simulations. Taking the full permutation of component labels of $\mathbf{z}^o$ and θ^o (inspired by Berkhof et al. (2003) and Marin and Robert (2008)), we thus propose a symmetrised version of a MAP proposal

$$\begin{aligned} q(\theta) &= \frac{1}{k!}\sum_{\sigma\in\mathfrak{S}_k}\pi_k(\theta|\sigma(\theta^o,\mathbf{z}^o),\mathbf{x}) \qquad (6)\\ &= \frac{1}{k!}\sum_{\sigma\in\mathfrak{S}_k}p(\boldsymbol{\lambda}|\sigma(\mathbf{z}^o))\prod_{i=1}^{k}p(\xi_i|\sigma(\boldsymbol{\xi}^o),\sigma(\mathbf{z}^o),\mathbf{x})\,. \end{aligned}$$

This proposal is equivalent to generating θ from $\pi_k(\theta|\theta^o,\mathbf{z}^o,\mathbf{x})$ and then operating a random permutation on the components of θ. The computational cost of producing $\omega(\theta)$ in (5), hence $\widehat{\mathfrak{E}}(k)$, is then multiplied by $k!$ under the provision that the support of (6) is sufficiently wide. If the tails of samples generated from (6) are deemed to be too narrow, as signalled by the effective sample size, additional selected (and thinned) simulations $\mathbf{z}^1,\ldots,\mathbf{z}^t$ taken from the Gibbs output can be included to make the proposal more robust.

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.

3.2 Dual importance sampling

A dual exploitation of the above Rao–Blackwellisation argument produces an alternative importance sampling proposal, based on MCMC draws $\{\theta^{(j)},\mathbf{z}^{(j)}\}_{j=1}^J$ from the unconstrained posterior distribution. The new importance function is expressed as

$$\begin{aligned} q(\theta) &= \frac{1}{Jk!}\sum_{j=1}^{J}\sum_{\sigma\in\mathfrak{S}_k}\pi_k(\theta|\sigma(\theta^{(j)},\mathbf{z}^{(j)}),\mathbf{x}) \qquad (7)\\ &= \frac{1}{Jk!}\sum_{j=1}^{J}\sum_{\sigma\in\mathfrak{S}_k}p(\boldsymbol{\lambda}|\sigma(\mathbf{z}^{(j)}))\prod_{i=1}^{k}p(\xi_i|\sigma(\boldsymbol{\xi}^{(j)}),\sigma(\mathbf{z}^{(j)}),\mathbf{x})\,. \end{aligned}$$

Here, $\pi_k(\theta|\sigma(\theta^{(j)},\mathbf{z}^{(j)}),\mathbf{x})$ is a product of full conditional densities on each parameter in a Gibbs sampler representation and $\{\theta^{(j)},\mathbf{z}^{(j)}\}_{j=1}^J$ is the original albeit not necessarily well-mixed simulation outcome. Label switching is imposed upon those J conditional densities through all $k!$ permutations and conversely the average of J well-selected conditional densities should approximate the posterior around any of the $k!$ symmetric modes of this posterior.

If we now assume that the component labels of the terms $\{\theta^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^{J}$ in (7) have not been permuted and that any label switching occurence has been removed from the simulations by a recentering method (Celeux et al. 2000), we denote the resulting transforms by $\{\varphi^{(j)}\}_{j=1}^{J}$. They can be interpreted as hyperparameters of q. The density (7) then satisfies

$$q(\theta) = \frac{1}{Jk!} \sum_{j=1}^{J} \sum_{i=1}^{k!} \pi(\theta|\sigma_i(\varphi^{(j)}), \mathbf{x}) \overset{\mathrm{ef}}{=} \frac{1}{k!} \sum_{i=1}^{k!} h_{\sigma_i}(\theta) \tag{8}$$

where $h_{\sigma_i}(\theta) = \dfrac{1}{J} \displaystyle\sum_{j=1}^{J} \pi(\theta|\sigma_i(\varphi^{(j)}), \mathbf{x})$. Each of the densities $h_{\sigma_1}, \cdots, h_{\sigma_{k!}}$ has a support– i.e., a domain where it takes non-negligible values– denoted by $S_{\sigma_1}, \cdots, S_{\sigma_{k!}}$ and $S_q = \bigcup_{i=1}^{k!} S_{\sigma_i}$. Note that an estimator using (8) is equivalent to an estimator using (7).

From a computational perspective, an artificial label switching step is necessary in computing $q(\theta)$ but not in generating a proposal θ from q. For arbitrary permutation representations $\sigma_m, \sigma_c, \sigma_i \in \mathfrak{S}_k = \{\sigma_1, \ldots, \sigma_{k!}\}$ acting on both θ and φ, the following holds for (7)

$$\pi(\sigma_c(\theta)|\sigma_i(\varphi), \mathbf{x}) = \pi(\sigma_m\sigma_c(\theta)|\sigma_m\sigma_i(\varphi), \mathbf{x})\,,$$

where $\sigma_m\sigma_c(\theta) = \sigma_m(\sigma_c(\theta))$. The full permutation representation set is invariant over an additional permutation representation σ_m (e.g., $\mathfrak{S}_k = \{\sigma_m\sigma_1, \cdots, \sigma_m\sigma_{k!}\}$), $q(\sigma_c(\theta))$ and $q(\sigma_m\sigma_c(\theta))$ are equal. Thus the standard estimator using q in (7) is equivalent (from a computational viewpoint) to an estimator such that (i) proposals are generated from a particular term $h_{\sigma_c}(\theta)$ of (8) and (ii) importance weights are computed according to (8). This makes a proposal generating step easier by ignoring label switching even though all the $h_\sigma(\theta)$'s need be evaluated to compute $q(\theta)$.

3.3 Importance function based on marginal posterior densities

Importance functions found in (4) and (8) have the same underlying motivation of a better approximation of the joint posterior distribution and the resulting estimate of (5) should therefore be more efficient. Both are designed to cover the $k!$ clusters, which are created by either symmetrizing the labels of hyperparameter set $\{\theta^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^{J}$ as in (8) or by randomly permuting the label of each $\{\theta^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^{J_1}$ as in (4). Once $k!$ clusters of hyperparameters are thus constructed, the corresponding conditional densities constitute clusters for q.

Consider $\kappa \in \{1, \ldots, k!\}$, a cluster index of q. Associating the cluster component function $q_\kappa(\cdot|\mathbf{x})$ with a support S_κ, the importance function q is expressed as

$$q(\theta|\mathbf{x}) = \sum_{\kappa=1}^{k!} p(\kappa) q_\kappa(\theta|\mathbf{x}) \tag{9}$$

where $p(\kappa)$ is the proportion of those conditional densities that are associated with the cluster κ and $\sum_{\kappa=1}^{k!} p(\kappa) = 1$. The importance function representation (8) is thus a special case of (9) with ($\kappa = 1, \ldots, k!$)

$$S_{\sigma_\kappa} = S_\kappa\,,\ h_{\sigma_\kappa}(\theta) = q_\kappa(\theta|\mathbf{x}) \ \text{ and } \ p(\kappa) = 1/k!\,.$$

By contrast, the density (4) does not achieve perfect symmetry, which means κ is not uniformly distributed, although $p(\kappa) \to 1/k!$ as $J_1 \to \infty$.

A marginal likelihood estimate using $q(\theta)$ as in (9) follows by a standard importance sampling identity

$$\begin{aligned}\mathfrak{E}(k) &= \int_{S_q} \frac{\pi(\theta)p_k(\mathbf{x}|\theta)}{q(\theta|\mathbf{x})}\left(\sum_{\kappa=1}^{k!} p(\kappa)q_\kappa(\theta|\mathbf{x})\right)\mathrm{d}\theta \\ &= \sum_{\kappa=1}^{k!}\int_{S_\kappa} \frac{\pi(\theta)p_k(\mathbf{x}|\theta)}{q(\theta|\mathbf{x})} p(\kappa)q_\kappa(\theta|\mathbf{x})\mathrm{d}\theta = \mathbb{E}_{p(\theta,\kappa)}[\omega(\theta)] \qquad (10)\end{aligned}$$

leading to

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

where $\omega(\theta) = \pi(\theta)p_k(\mathbf{x}|\theta)/q(\theta|\mathbf{x})$, namely a weighted sum of integrals over the S_κ's ($\kappa = 1, \ldots, k!$).

Due to the perfect symmetry in the clusters of (8), the integrals of ωq_κ with respect to θ over S_κ for $\kappa = 1, \cdots, k!$ are equal. Given an arbitrary cluster, κ^o, the evidence is

$$\begin{aligned}\mathfrak{E}(k) &= \sum_{\kappa=1}^{k!} p(\kappa)\left(\int_{S_\kappa}\omega(\theta)q_\kappa(\theta|\mathbf{x})\mathrm{d}\theta\right) \\ &= \int_{S_{\kappa^o}}\omega(\theta)q_{\kappa^o}(\theta|\mathbf{x})\mathrm{d}\theta = \mathbb{E}_{q_{\kappa^o}(\theta|\mathbf{x})}[\omega(\theta)]\ . \qquad (11)\end{aligned}$$

Note that the corresponding estimator (Monte Carlo approximation based on T draws) for the above is exactly in the same form to the estimator for (10).

Both (10) and (11) are thus importance sampling estimators using (4) and (8) respectively. Hence standard convergence result hold: by the Law of Large Numbers, both estimates a.s. converge to $\mathfrak{E}(k)$, and the Central Limit theorem also holds

$$\sqrt{T}\left\{\frac{1}{T}\sum_{t=1}^{T}\omega(\theta^{(t)}) - \mathfrak{E}(k)\right\}\underset{T\to\infty}{\longrightarrow}\mathcal{N}(0,V_1)\,,\ \sqrt{T}\left\{\frac{1}{T}\sum_{t=1}^{T}\omega(\theta^{(t)}) - \mathfrak{E}(k)\right\}\underset{T\to\infty}{\longrightarrow}\mathcal{N}(0,V_2)$$

where $V_1 = \mathrm{var}_{p(\theta,\kappa|\mathbf{x})}(\omega(\theta))$ and $V_2 = \mathrm{var}_{q_{\kappa^o}(\theta|\mathbf{x})}(\omega(\theta))$. The perfect symmetry in the clusters of (8) does not guarantee a better efficiency in estimation and those variances are rather highly related to how well the importance functions approximate the joint posterior distribution. If $J_1 = Jk!$ and both importance functions provide a good approximation, $V_1 \approx V_2$ is expected.

4 Importance function approximation

Both estimators (10) and (11) suffer from massive computational demands when k is large. In this section, we show how to approximate (7) and increase the computational efficiency (i.e., computing time) as a result.

It was shown in Section 3.2 that q as in (7) is invariant under a permutation of the labels of θ and that proposals can be generated from a particular term $h_{\sigma_c}(\theta)$ of (8) without any loss of statistical efficiency. Given $(\theta^{(1)}, \ldots, \theta^{(T)}) \sim h_{\sigma_c}(\theta)$, it is natural to consider whether or not all terms in $\{h_{\sigma_1}(\theta^{(t)}), \ldots, h_{\sigma_{k!}}(\theta^{(t)})\}$ are different from zero for $t = 1, \ldots, T$. In the case some are not, it is obviously computationally relevant to determine which ones among them are likely to be insignificant (i.e., almost zero). This perspective motivates the following section.

4.1 Dual importance sampling using an approximation

Given a proposal θ generated from a particular $h_{\sigma_c}(\theta)$, $\theta \in S_{\sigma_c}$, the importance function at θ is an average of all $h_\sigma(\theta)$'s and the relative contribution of each term is

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

If $\eta_{\sigma_i}(\theta)$ is close to zero, $h_{\sigma_i}(\theta)$ is negligible within $q(\theta)$ and on the opposite $\eta_{\sigma_i}(\theta) \approx 1$ indicates a high contribution of $h_{\sigma_i}(\theta)$. The expected relative contribution of $h_{\sigma_i}(\theta)$

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

is estimated by

$$\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_i}(\theta)] = \frac{1}{M} \sum_{l=1}^{M} \eta_{\sigma_i}(\theta^{(l)}) \ , \qquad \theta^{(l)} \sim h_{\sigma_c}(\theta) \ . \tag{12}$$

After an appropriate permutation of the indices, we obtain that $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}(\theta)] \geq \cdots \geq \widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}(\theta)]$, namely that the corresponding $h_{\sigma_1}, \cdots, h_{\sigma_{k!}}$ are in decreasing order of expected contributions. The importance function $q(\theta)$ can then be approximated by using only the n most important h_σ's ($1 \leq n \leq k!$), leading to the approximation

$$\tilde{q}_n(\theta) = \frac{1}{k!} \sum_{i=1}^{n} h_{\sigma_i}(\theta) \ , \tag{13}$$

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

$$\hat{\phi}_n = \frac{1}{M} \sum_{l=1}^{M} \left| \tilde{q}_n(\theta^{(l)}) - q(\theta^{(l)}) \right| \ , \qquad \theta^{(l)} \sim h_{\sigma_c}(\theta) \ . \tag{14}$$

When this mean absolute difference is below a certain threshold, τ, $\tilde{q}_n$ is considered to be an appropriate approximation for q. We define the corresponding approximate set $\mathfrak{A}(k) \subseteq \mathfrak{S}_k$ to be made of $\{\sigma_1, \cdots, \sigma_n\}$, n being defined as the smallest size that satisfies the condition $\widehat{\phi}_n < \tau$. With this truncation, the computational efficiency obviously improves.

Note that the set $\mathfrak{A}(k)$ is determined under the assumption that all proposals $(\theta^{(t)})$ are generated from h_{σ_c} since the quality of the approximation is only guaranteed under this assumption. Due to the perfect symmetry of $q(\theta)$ over the $k!$ permutations, the choice of σ_c is obviously irrelevant for the computational gains. The evidence estimate using such an approximation is detailed in the following algorithm:

```
Algorithm 1: Dual importance sampling algorithm with approximation

1 Randomly select $\{\mathbf{z}^{(j)}, \theta^{(j)}\}_{j=1}^J$ from Gibbs sample and remove label switching by an
    appropriate method. Then, construct $q(\theta)$ as in (8).
2 Derive the corresponding term $h_{\sigma_c}(\theta)$ and generate particles $\{\theta^{(t)}\}_{t=1}^T \sim h_{\sigma_c}(\theta)$.
3 Construct an approximation, $\tilde{q}(\theta)$, using the first $M$ terms in $\{\theta^{(t)}\}_{t=1}^T$:
    3.1 Compute $(h_{\sigma_1}(\theta^{(t)}), \ldots, h_{\sigma_{k!}}(\theta^{(t)}), \eta_{\sigma_1}(\theta^{(t)}), \ldots, \eta_{\sigma_{k!}}(\theta^{(t)}))$ for $t = 1, \ldots, M$
        and $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}(\theta)], \cdots, \widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}(\theta)]$ as in (12).
    3.2 Reorder the $\sigma$'s so that $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}(\theta)] \geq \cdots \geq \widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}(\theta)]$.
    3.3 Initialise $n = 1$ and compute $\tilde{q}_n(\theta^{(1)}), \cdots, \tilde{q}_n(\theta^{(M)})$ as in (13) and $\widehat{\phi}_n$ as in
        (14). If $\widehat{\phi}_{n=1} < \tau$, go to Step 4. Otherwise increase $n = n + 1$ and update
        $\tilde{q}_n(\theta)$ and $\widehat{\phi}_n(\theta)$ until $\widehat{\phi}_n < \tau$.
4 Calculate $\tilde{q}_n(\theta^{(M+1)}), \ldots, \tilde{q}_n(\theta^{(T)})$ and replace $q(\theta^{(1)}), \ldots, q(\theta^{(T)})$ with
    $\tilde{q}(\theta^{(1)}), \ldots, \tilde{q}(\theta^{(T)})$ in (5) to estimate $\widehat{\mathfrak{E}}$.
```

In Step 1., we used the method by Jasra et al. (2005), even though alternatives implemented in the label.switching package of Papastamoulis and Iliopoulos (2010) or in Rodriguez and Walker (2014) could be implemented as well. The total number of h values that are computed is $Tk!$ in the standard dual importance sampling scheme but decreases to $(Mk!) + |\mathfrak{A}(k)|(T - M)$ when using $\tilde{q}(\theta)$. The relative gain in the total number of terms is thus

$$\Delta(\mathfrak{A}(k)) = \frac{(Mk!) + |\mathfrak{A}(k)|(T - M)}{Tk!} = \frac{M}{T}\left(1 - \frac{|\mathfrak{A}(k)|}{k!}\right) + \frac{|\mathfrak{A}(k)|}{k!}\ . \tag{15}$$

The gain will thus depend on how small $|\mathfrak{A}(k)|$ is, when compared with $k!$, hence ultimately on the acceptable mean absolute difference τ.

5 Simulation study

Two simulated mixture datasets and two real datasets are used to examine the performance of seven marginal likelihood estimators. The simulated datasets, D_1 and D_2, are;

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

where $N(5, 2^2)$ denotes a normal distribution with a mean of 5 and a standard deviation of 2. Two real datasets, called galaxy and fishery datasets respectively, are shown in Figure 1. They have been frequently used in the literature as benchmarks (see, e.g. Chib 1995; Frühwirth-Schnatter 2006; Jasra et al. 2005; Richardson and Green 1997; Stephens 2000b).

Gaussian and Dirichlet priors are used for the means $\{\mu_i\}_{i=1}^k$ and proportions $\boldsymbol{\lambda}$,

$$\{\mu_i\}_{i=1}^k \sim N(0, 10^2) \quad \text{and} \quad (\lambda_1, \ldots, \lambda_k) \sim \text{Dir}(1, \ldots, 1) \ .$$

For the variance parameters $\{\sigma_i^2\}_{i=1}^k$, inverse Gamma distributions with two sets of hyperparameters, $IG(2, 3)$ and $IG(2, 15)$, are considered. With the second calibration, label switching naturally occurred in Gibbs sequences in our simulation experiments. Removing the first 5000 Gibbs simulations as burn-ins, 10^4 Gibbs simulations are used to approximate $\mathfrak{E}(k)$.

Firstly, a sensitivity analysis is conducted about the expected relative contribution of h_{σ_i} to $q(\theta)$ with respect to M. Then we set the values for both M and τ. In Section 5.2, the performance of seven estimators for $\mathfrak{E}(k)$ are compared through a large simulation study, which confirms that the asymptotic variance of $\widehat{\mathfrak{E}}(k)$ based on (7) is smaller than when based on (4).

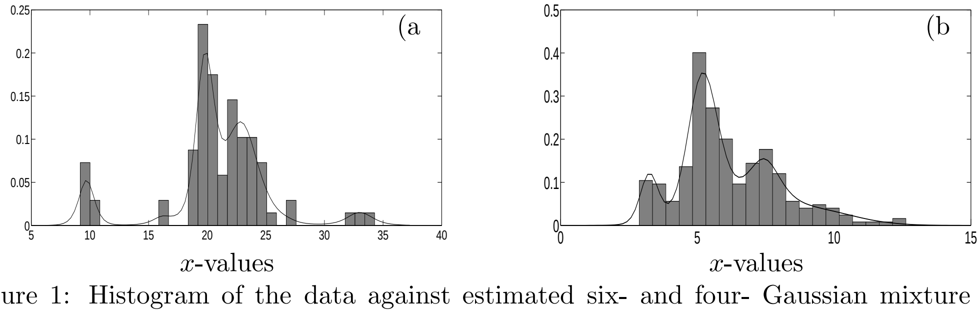

Figure 1: Histogram of the data against estimated six- and four- Gaussian mixture densities (solid line) for (a) the Galaxy dataset and (b) the fishery dataset, respectively.

5.1 Determining M and τ

The approximation set is constructed in two steps. First, we compute $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1(\theta)}]$, …, $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}(\theta)}]$, based on reduced samples of size M as in (12). Second, we derive which terms are negligible when compared with the threshold τ. In our experiments, we chose τ conservatively so that all zero terms are identified. In MatLab, 10^{-324} is rounded down to 0 thus $\tau = 10^{-324}$ was chosen for the following simulation studies.

The expected relative contribution measures for D_1 and D_2 are given in Tables 1 and 2, respectively. For $J = 10^2$ initial Gibbs simulations, significantly contributing clusters are easily identified by $\{\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_i}(\theta)]\}_{i=1}^{k!}$ and both $|\mathfrak{A}(k)|$ and $\widehat{\phi}$ are relatively stable against M. Under a natural lack of label switching, $q(\theta)$ seems to be well approximated using only $h_{\sigma_1}(\theta)$, as seen in Table 1. Even when some label switching occurs in a Gibbs sequence corresponding to a Gaussian mixture model with three components, only two terms, $h_{\sigma_1}(\theta)$ and $h_{\sigma_2}(\theta)$, significantly contribute to $q(\theta)$, as seen in Table 2. For the subsequent analyses in this paper, we chose $J = 10^2$, $M = 10^3$ and $\tau = 10^{-324}$.

M	$\{\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_i}(\theta)]\}_{i=1}^{k!}$	$\lvert\mathfrak{A}(k)\rvert$	$\widehat{\phi}$
10^2	$[1, 1.89 \times 10^{-102}]$	1	0
10^3	$[1, 5.25 \times 10^{-90}]$	1	0
10^4	$[1, 4.62 \times 10^{-91}]$	1	0
10^5	$[1, 3.56 \times 10^{-80}]$	1	0

Table 1: Estimates for $\{\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_i}]\}_{i=1}^{k!}$, $|\mathfrak{A}(k)|$ and $\widehat{\phi}$ against M for D_1 ($k = 2$). The prior for a variance parameter is $IG(2, 3)$. Note that due to rounding errors, the sum of the contribution ratios does not equal one.

M	$\{\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_i}]\}_{i=1}^{k!}$	$\lvert\mathfrak{A}(k)\rvert$	$\widehat{\phi}$
10^2	$[3.56 \times 10^{-16}, 9.53 \times 10^{-160}, 5.05 \times 10^{-55}, 8.27 \times 10^{-144}, 1.0, 4.64 \times 10^{-65}]$	2	0
10^3	$[1.22 \times 10^{-8}, 1.11 \times 10^{-144}, 3.01 \times 10^{-49}, 3.08 \times 10^{-125}, 1.0, 2.27 \times 10^{-53}]$	2	0
10^4	$[2.03 \times 10^{-8}, 8.31 \times 10^{-136}, 1.76 \times 10^{-43}, 2.61 \times 10^{-95}, 1.0, 4.87 \times 10^{-49}]$	2	0
10^5	$[1.04 \times 10^{-5}, 1.56 \times 10^{-122}, 1.51 \times 10^{-44}, 4.30 \times 10^{-87}, 1.0, 2.27 \times 10^{-39}]$	2	0

Table 2: Estimates for $\{\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_i]\}_{i=1}^{k!}$, $|\mathfrak{A}(k)|$ and $\widehat{\phi}$ with respect to M for D_2 ($k = 3$). The prior for a variance parameter is $IG(2, 15)$. Note that due to rounding errors, the sum of the contribution ratios does not equal one.

5.2 Simulation results

The following seven marginal likelihood estimators using an equal number of proposals are compared;

$\widehat{\mathfrak{E}}^*_{Ch}$ **:** Chib's method (2) using $T = 10^4$ samples and multiplying by $k!$ to compensate for a lack of label switching;

$\widehat{\mathfrak{E}}_{Ch}$ **:** Chib's method with density estimate (2), using $T = 10^4$ randomly permuted Gibbs samples;

$\widehat{\mathfrak{E}}_{IS}$ **:** Importance sampling using q as in (6), with a maximum likelihood estimate for $z^o_1, \dots, z^o_n$ and $T = 10^4$ particles;

$\widehat{\mathfrak{E}}_{DS}$ **:** Dual importance sampling using q as in (7), with $T = 10^4$ particles and $J = 100$ Gibbs samples in $q(\theta)$;

$\widehat{\mathfrak{E}}^A_{DS}$ **:** Dual importance sampling using an approximation as in (13), with $T = 10^4$ particles, $J = 100$ and $M = 10^3$;

$\widehat{\mathfrak{E}}_{J_1}$ **:** Importance sampling using q as in (4), with $T = 10^4$ particles. When $k < 6$, $J_1 = 100k!$ and otherwise $J_1 = 5000$;

$\widehat{\mathfrak{E}}_{BS}$ **:** Bridge sampling (3), using $M_1 = M_2 = 6 \times 10^3$ samples and $q(\theta)$ as in (4) via 10 iterations. For q, it is set as $J_1 = 4000$ and label switching is imposed in hyperparameters $\{\theta^{(j)}, z^{(j)}\}^{J_1}_{j=1}$.

The marginal likelihood estimates (in log scales) and the effective sample size (ESS) ratios, $R = \mathrm{ESS}/T$, are summarized in Figures 2 and 3 by boxplots, based on 50 replicates. Subscripts of $\widehat{\mathfrak{E}}$ and R denote the estimating technique. Note that a modified ESS, provided by equation (35) in Doucet et al. (2000), is used here for numerical stability. All estimators are based on 10^4 proposals, as in Table 3, where summing up the second and third columns leads to a fixed total number of function evaluations. Within our setup, $\widehat{\mathfrak{E}}_{IS}$ is the least demanding in terms of computational workload while the remaining importance estimators require the same computing time, except for $\widehat{\mathfrak{E}}^A_{DS}$.

Simulated mixture dataset

Mixture models of two and three components are fitted to D_1 and D_2 respectively. Regardless of the presence or not of label switching in the resulting Gibbs sequences, all estimates based on importance sampling except $\widehat{\mathfrak{E}}_{IS}$ coincide with $\widehat{\mathfrak{E}}_{Ch}$, albeit with smaller Monte Carlo variations as seen in Figures 2 and 3. When a suitable approximate for $q(\theta)$ is used for the dual importance sampling, no significant difference in the estimates $\log(\widehat{\mathfrak{E}}(k))$ and in the effective sample sizes are observed. The mean sizes of $\mathfrak{A}(k)$ in Table 4 are always smaller than $k!$ and it shows that $\mathfrak{E}(k)$ can be estimated with a lesser computational workload. When posterior modes are very well separated (no

Estimate	Number of posterior evaluations	Number of marginal posterior density evaluations in q	Number of proposals
$\widehat{\mathfrak{E}}_{IS}$	T	$Tk!$	T
$\widehat{\mathfrak{E}}_{DS}$	T	$TJk!$	T
$\widehat{\mathfrak{E}}^{A}_{DS}$	T	$(M+(T-M)\lvert\mathfrak{A}(k)\rvert/k!)Jk!$	T
$\widehat{\mathfrak{E}}_{J_1}$	T	TJ_1	T
$\widehat{\mathfrak{E}}_{BS}$	M_1	$(M_1+M_2)J_1$	M_1+M_2

Table 3: Computation steps required by different evidence estimation approaches. Note that the required computation for $\widehat{\mathfrak{E}}_{BS}$ is given per iteration.

natural label switching ever present in Gibb sequences), the number of evaluations in q is reduced almost by the maximal factor of $1/k!$. In Table 5, the least computational demand is observed for the chib's methods while the bridge sampling costs more than 100 times. When $\mathfrak{A}(k) < k!$, some reduction in CPU time for $\widehat{\mathfrak{E}}(k)^{A}_{DS}$ is observed due to ignoring zero function evaluation.

Disagreement in the values of $\widehat{\mathfrak{E}}_{IS}$ versus $\widehat{\mathfrak{E}}_{Ch}$ shows that an importance function may fail to properly approximate $p_k(\mathbf{x}|\theta)\pi(\theta)$, resulting in an unreliable estimate with large variation. Significantly small effective sample sizes (i.e., very small values for R_{IS}) back this observation. In our simulation experiments, we observed that $\widehat{\mathfrak{E}}_{BS}$ is correctly calibrated for a large value of J_1 (i.e., a large number of conditional densities in q). When label switching naturally occurs, as in the Gibbs sequence under the variance prior $IG(2,15)$, $\widehat{\mathfrak{E}}^{*}_{Ch}$ disagrees with the other estimates, see Figure 3. Unsurprisingly, this indicates that the simplistic correction through a multiplication by $k!$ is of no use, as reported in Neal (1999), Frühwirth-Schnatter (2006) and Marin and Robert (2008).

D	k	$k!$	$\lvert\mathfrak{A}_1(k)\rvert$	$\Delta(\mathfrak{A}_1)$	$\lvert\mathfrak{A}_2(k)\rvert$	$\Delta(\mathfrak{A}_2)$
D_1	2	2	1.00 (0.00)	0.55 (2.26×10^{-16})	1.73 (0.45)	0.88 (0.20)
D_2	3	6	1.02 (0.14)	0.25 (0.02)	2.18 (0.60)	0.43 (0.09)

Table 4: Mean and standard deviation *(values in brackets)* estimates for the approximation set size, $|\mathfrak{A}(k)|$, and the reduction rate of a number of evaluated h-terms, $\Delta(\mathfrak{A})$, as in (15) for D_1 and D_2. Subscripts 1 and 2 indicate the results using the priors $\sigma^2 \sim IG(2,3)$ and $\sigma^2 \sim IG(2,15)$, respectively.

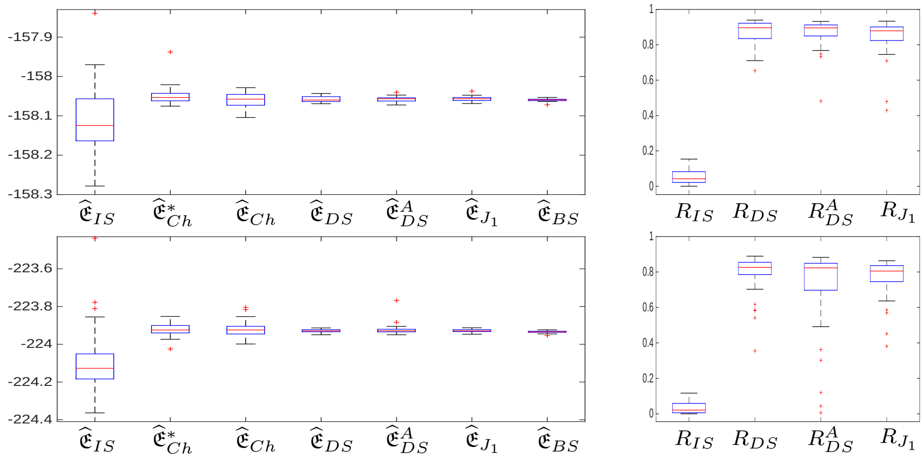

Figure 2: Boxplots of evidence estimates in log scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with two and three Gaussian components are fitted to *(top)* D_1 and *(bottom)* D_2, respectively. The prior for $\{\sigma_i^2\}_{i=1}^k$ is $IG(2,3)$ and label switching did not occur in Gibbs samples. One outlier of $\widehat{\mathfrak{E}}_{IS}$ in the top-left panel is discarded.

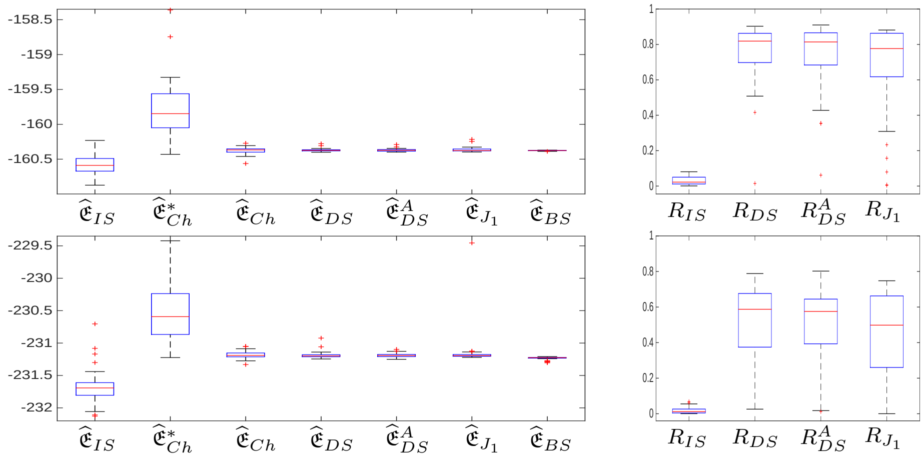

Figure 3: Boxplots of evidence estimates in log scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with two and three Gaussian components are fitted to *(top)* D_1 and *(bottom)* D_2, respectively. The prior for $\{\sigma_i^2\}_{i=1}^k$ is $IG(2,15)$ and label switching naturally occurred in Gibbs samples. Two outliers for $\widehat{\mathfrak{E}}^*_{Ch}$ in the top-left panel are discarded.

Galaxy and fishery dataset

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

$$\begin{array}{rcl} \mu_1,\ldots,\mu_k & \sim & N(\bar{\mathbf{x}}, r^2/4) \\ \sigma_1^2,\ldots,\sigma_k^2 & \sim & IG(2,\beta) \\ \beta & \sim & G(0.2, 10/r^2) \\ \lambda_1,\ldots,\lambda_k & \sim & \text{Dirichlet}(1,\ldots,1) \end{array}$$

Estimator	D_1		D_2	
	CPU$_1$	CPU$_2$	CPU$_1$	CPU$_2$
$\widehat{\mathfrak{E}}^*_{Ch}$	0.80	0.76	1.17	1.39
$\widehat{\mathfrak{E}}_{Ch}$	0.79	0.81	1.32	1.36
$\widehat{\mathfrak{E}}_{IS}$	2.54	2.65	3.35	3.35
$\widehat{\mathfrak{E}}_{DS}$	3.07	2.96	6.12	6.02
$\widehat{\mathfrak{E}}^A_{DS}$	2.87	3.07	3.77	5.47
$\widehat{\mathfrak{E}}_{J_1}$	2.42	3.34	6.02	6.88
$\widehat{\mathfrak{E}}_{BS}$	1.18×10^3	1.19×10^3	2.78×10^3	3.37×10^3

Table 5: Elapsed CPU time in seconds for evidences approximation of mixture models for D_1 and D_2. Subscripts 1 and 2 of CPU indicate the results using the priors $\sigma^2 \sim IG(2,3)$ and $\sigma^2 \sim IG(2,15)$, respectively.

where $\bar{\mathbf{x}}$ and r are the median and the range of $\mathbf{x}$, respectively. Normal mixture models are fitted to both datasets and estimates of $\log(\mathfrak{E}(k))$ and R are summarized in Figures 4 and 5. In general, a similar behaviour of $\log(\widehat{\mathfrak{E}}(k))$ and R between the methods is observed. For all cases, the dual importance sampling schemes ($\widehat{\mathfrak{E}}_{DS}$ and $\widehat{\mathfrak{E}}^A_{DS}$) and $\widehat{\mathfrak{E}}_{J_1}$ agree with Chib's approach ($\widehat{\mathfrak{E}}_{Ch}$). Unless modes of the joint posterior distributions are clearly separated (e.g., $|\overline{\mathfrak{A}(k)}| \approx 1$), $\log(\widehat{\mathfrak{E}}^*_{Ch})$ is biased due to an improper permutation correction. When a poor $q(\theta)$ is used for importance sampling, inaccurate approximations result and the range of $\widehat{\mathfrak{E}}_{IS}$ estimates is much off from the other estimates.

Symptoms of the "curse of dimensionality" can be observed. As k increases, the effective sample size decreases exponentially fast and the variation in the estimates increases. Given the complex shape of the posterior distribution, the support common to $q(\theta)$ and $\pi^*_k(\theta|\mathbf{z})$ gets progressively smaller and $\widehat{\mathfrak{E}}_{BS}$ becomes less accurate, as shown in both figures. When $k = 6$, the variation in the values of $\widehat{\mathfrak{E}}_{Ch}$ is much larger than those of the estimate by dual importance sampling. When $J_1 \ll Jk!$, q does not provide a good approximation of the joint posterior and $\log(\mathfrak{E}_{J_1})$ is then biased. Due to a fast increase of $k!$, fast increasing in CPU times is seen for all estimators in Table 7.

The reduction in the number of evaluated terms used to approximate $\widehat{\mathfrak{E}}(k)$ varies case by case, as shown in Table 6. When $k = 4$ and $k = 6$, components of the posterior distribution for the galaxy data tend to have long flat tails and thus have higher chance to overlap each other. Consequently, the workload reduction is of lesser magnitude than for a model with a smaller number of components. Provided that some functions are neglected for $\widehat{\mathfrak{E}}^A_{DS}$, there is some gain in computational efficiency as can be seen in Table 6.

k	$k!$	$\lvert\mathfrak{A}(k)\rvert$	$\Delta(\mathfrak{A})$
3	6	1.00 (0.00)	0.25 (0.00)
4	24	2.10 (0.76)	0.18 (0.03)

(a) Fishery data

k	$k!$	$\lvert\mathfrak{A}(k)\rvert$	$\Delta(\mathfrak{A})$
3	6	1.06 (0.24)	0.26 (0.04)
4	24	13.34 (5.35)	0.60 (0.20)
6	720	176.78 (75.31)	0.32 (0.09)

(b) Galaxy data

Table 6: Mean and standard deviation *(values in brackets)* of approximate set sizes, $|\mathfrak{A}(k)|$, and the reduction rate of a number of evaluated h-terms $\Delta(\mathfrak{A})$ as in (15) for (a) fishery and (b) galaxy datasets.

Estimator	Fishery data		Galaxy data		
	$k=3$	$k=4$	$k=3$	$k=4$	$k=6$
$\widehat{\mathfrak{E}}^{*}_{Ch}$	1.71	1.71	1.20	1.88	2.89
$\widehat{\mathfrak{E}}_{Ch}$	1.40	2.30	1.56	2.18	26.86
$\widehat{\mathfrak{E}}_{IS}$	12.23	14.60	13.47	14.83	48.74
$\widehat{\mathfrak{E}}_{DS}$	27.75	86.98	27.00	85.27	3.10×10^3
$\widehat{\mathfrak{E}}^{A}_{DS}$	18.14	30.45	18.28	52.49	1.33×10^3
$\widehat{\mathfrak{E}}_{J_1}$	28.19	90.11	26.75	87.19	244.10
$\widehat{\mathfrak{E}}_{BS}$	4.92×10^3	6.71×10^3	4.21×10^3	3.14×10^3	7.32×10^3

Table 7: Elapsed CPU time in seconds for evidences approximation of mixture models for fishery and galaxy datasets.

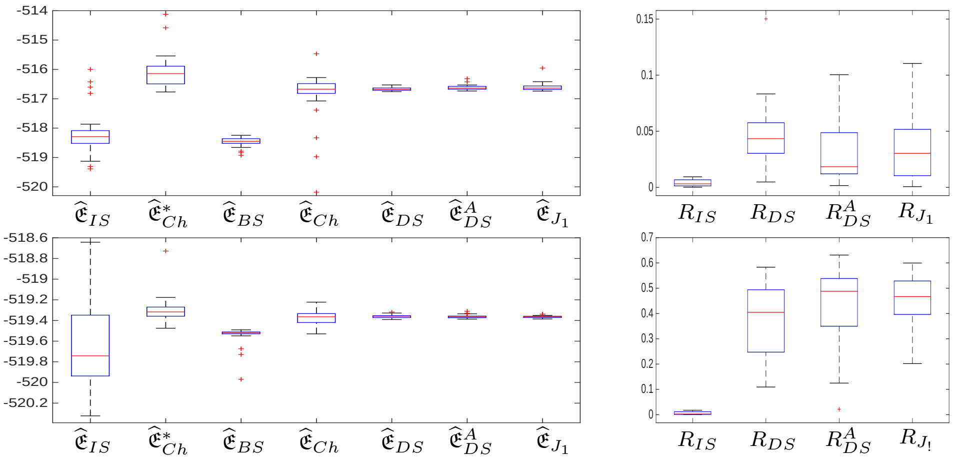

Figure 4: Boxplots of evidence estimates in log scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with *(top)* three and *(bottom)* four Gaussian components are fitted to the fishery dataset. Two outliers of $\widehat{\mathfrak{E}}_{Ch}$ in the top-left panel are discarded.

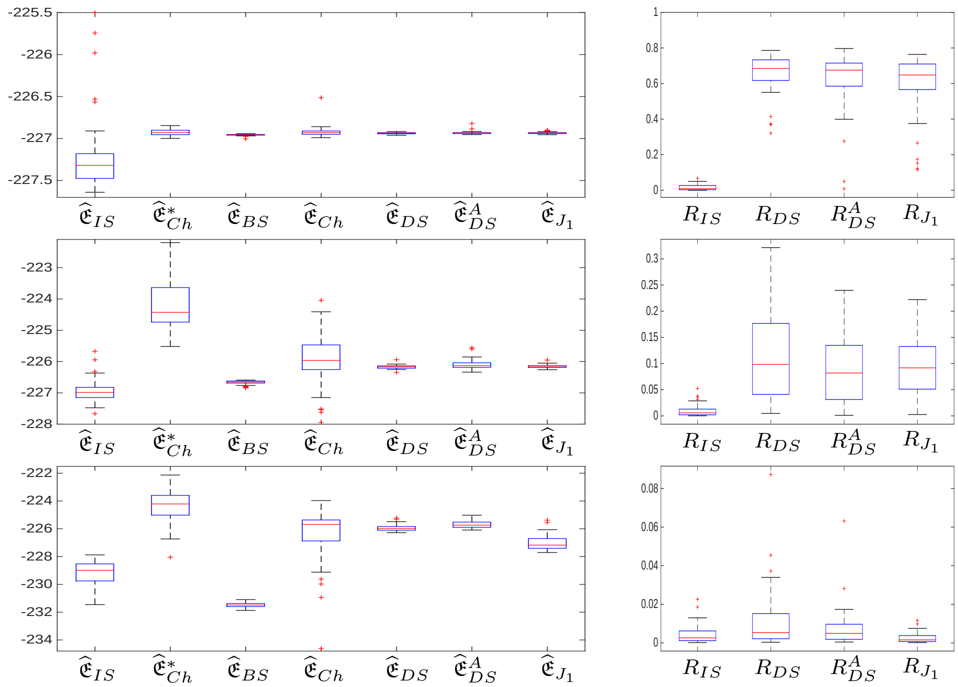

Figure 5: Boxplots of evidence estimates in log scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with *(top)* three, *(middle)* four, and *(bottom)* six Gaussian components are fitted to the galaxy dataset. One outlier of $\widehat{\mathfrak{E}}_{Ch}$ in the top-left panel is discarded.

6 Discussion

This paper considered evidence approximations by importance sampling for mixture models and re-evaluated some of the known challenges resulting from high multimodality in the posterior density. Importance sampling requires that the support of an importance function encompasses the support of the posterior density to perform properly. In the specific case on mixture models, missing some of the invariance under permutation function is likely to produce an unsuitable support hence, a poor estimate of the evidence.

In our study, exchangeable priors are used, which implies that the posterior and marginal posterior densities exhibit $k!$ symmetrical terms. Two marginal likelihood estimators are proposed here and tested against other existing estimators. The first approach exploits the permutation invariance of $\pi(\cdot|\mathbf{x}, \mathbf{z}^o)$ with a pointwise MLE, $\mathbf{z}^o$, to create an importance function. However, due to a poor resulting support, this approach performs quite poorly in our simulation studies. Another poor estimate is derived from Chib's method when the invariance by permutation is not reproduced in the sample (Neal 2001).

A second importance function is constructed by double Rao–Blackwellisation, hence the denomination of *dual importance sampling*. We demonstrate both methodologically and practically that this solution fits the demands of mixture estimation. Moreover, introducing a suitable and implementable approximation scheme, we show how to avoid the exponential increase in k of the computational workload. The idea at the core of this approximation is to bypass negligible elements in the approximation thanks to the perfect symmetry of the posterior density. When posterior modes are well-separated, the gain is of a larger magnitude than when those modes strongly overlap.

Borrowing from the original approach in Chib (1996), dual importance sampling can be extended to cases when conditional Gibbs sampling densities are not available in closed form. However, this solution suffers from the curse of dimensionality, just like any other importance sampling estimator.

Alternative evidence approximation techniques could be considered for this problem, as exemplified in Friel and Wyse (2012). For instance, *ensemble Monte Carlo* samples from local ensembles that are extensions or compositions of the original, e.g., using parallel tempering Monte Carlo methods. Extending this idea, Bayes factor approximations were proposed using annealed importance sampling (Neal 2001) and power posteriors (Friel and Pettitt 2008). Further investigation is needed to characterize the performances of those alternative solutions in the setting of mixture models and label switching.

References

Antoniak, C. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 2: 1152–1174.

Ardia, D., Baştürk, N., Hoogerhieide, L., and van Dijik, H. K. (2012). "A comparative study of Mnote Carlo methods for efficient evaluation of marginal likelihood." *Computational Statistics and Data Analysis*, 56: 3398–3414.

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

Carlin, B. and Chib, S. (1995). "Bayesian model choice through Markov chain Monte Carlo." *J. Royal Statist. Society Series B*, 57(3): 473–484.

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

Chen, M.-H., Shao, Q. M., and Ibrahim, J. G. (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. P. (2010). “Properties of Nested sampling.” *Biometrika*, 97: 741–755.

Congdon, P. (2006). “Bayesian model choice based on Monte Carlo estimates of posterior model probabilities.” *Computational Statistics and Data Analysis*, 50: 346–357.

DiCiccio, A. P., Kass, R. E., Raftery, A., and Wasserman, L. (1997). “Computing Bayes factors by combining simulation and asymptotic approximations.” *Journal of the American Statistical Association*, 92: 903–915.

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

Doucet, A., Godsill, S., and Andrieu, C. (2000). “On sequential Mnote Carlo sampling methods for Bayesian filtering.” *Statistics and Computing*, 10: 197–208.

Escobar, M. and West, M. (1995). “Bayesian density estimation and inference using mixtures.” *Journal of the American Statistical Association*, 90(430): 577–588.

Friel, N. and Pettitt, A. N. (2008). “Marginal likelihood estimation via power posteriors.” *Journal of the Royal Statistical Society, Series B*, 70: 589–607.

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 for 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.” *Journal of Econometrics*, 7: 143–167.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Series in Statistics, 1 edition.

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

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

Geweke, J. (2012). “Interpretation and inference in mixture models: simple MCMC works.” *Computational Statistics and Data Analysis*, 51: 3529–3550.

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. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, New York.

— (2010a). "Importance sampling methods for Bayesian discrimination between embedded models." In Chen, M.-H., Dey, D., Müller, P., Sun, D., and Ye, K. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*. Springer-Verlag, New York.

— (2010b). "On resolving the Savage–Dickey paradox." *Electron. J. Statist.*, 4: 643–654.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). "Bayesian modelling and inference on mixtures of distributions." In Rao, C. and Dey, D. (eds.), *Handbook of Statistics*, volume 25. Springer-Verlag, New York.

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

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

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

Papastamoulis, P. (2013). *label.switching: Relabelling MCMC outputs of mixture models*. R package version 1.2.
URL `http://CRAN.R-project.org/package=label.switching`

Papastamoulis, P. and Iliopoulos, G., G. (2010). "An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions." *Journal of Computational and Graphical Statistics*, 19(2): 313–331.

Papastamoulis, P. and Roberts, G. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95: 315–321.

Perrakie, K., Ntzoufras, I., and Tsionas, E. G. (2014). "On the use of marginal posteriors in marginal likelihood estimation via importance sampling." *Computational Statistics and Data Analysis*, 77: 54–69.

Raftery, A., Newton, M., Satagopan, J., and Krivitsky, P. (2006). "Estimating the integrated likelihood via posterior simulation using the harmonic mean identity." Technical Report 499, University of Washington, Department of Statistics.

Rasmussen, C. E. (2000). "The Infinite Gaussian Mixture Model." In *In Advances in Neural Information Processing Systems 12*, 554–560. MIT Press.

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 Marin, J.-M. (2008). "On some difficulties with a posterior probability approximation technique." *Bayesian Analysis*, 3(2): 427–442.

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

Rodriguez, C. and Walker, S. (2014). "Label switching in Bayesian mixture models:Deterministic relabeling strategies." *Journal of Computational and Graphical Statistics*, 21(1): 23–45.

Rubin, D. B. (1987). "Comment on "The calculation of posterior distributions by data augmentation" by M. A. Tanner and W. H. Wong." *Journal of the American Statistical Association*, 82: 543–546.

— (1988). "Using the SIR algorithm to simulate posterior distributions." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics, 3*, 395–402. Oxford University Press.

Satagopan, J., Newton, M., and Raftery, A. (2000). "Easy Estimation of Normalizing Constants and Bayes Factors from Posterior Simulation: Stabilizing the Harmonic Mean Estimator." Technical Report 1028, University of Wisconsin-Madison, Department of Statistics.

Scott, S. L. (2002). "Bayesian methods for hidden Markov models: recursive computing in the 21st Century." *Journal of the American Statistical Association*, 97: 337–351.

Servidea, J. D. (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. (2000a). “Bayesian Analysis of Mixture Models with an Unknown Number of Components - An Alternative to Reversible Jump Methods.” *The Annals of Statistics*, 28(1): 40–74.

— (2000b). “Dealing with label switching in mixture models.” *Journal of Royal Statistical Society, Series B*, 62: 795–809.

Tierney, L. and Kadane, J. (1986). “Accurate approximations for posterior moments and marginal densities.” *Journal of the American Statistical Association*, 81: 82–86.

Verdinelli, I. and Wasserman, L. (1995). “Computing Bayes factors using a generalization of the Savage–Dickey density ratio.” *Journal of the American Statistical Association*, 90: 614–618.

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

Acknowledgments

We are most grateful to the Editorial Team of Bayesian Analysis for their helpful suggestions and their support towards this revision. Financial support by CEREMADE, Université Paris-Dauphine and Auckland University of Technology for a visit of Jeong Eun Lee is most appreciated. Christian Robert is partially supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2010–2015 ANR-11-BS01-0010 grant “Calibration” and by a 2010–2015 Institut Universitaire de France senior chair.