Bayesian Analysis (2015) **TBA**, Number TBA, pp. 1–25

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 maximum likelihood estimate (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 less computational workload.

Keywords: model evidence, importance sampling, mixture models, marginal likelihood.

1 Introduction

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

$$X_j|\boldsymbol{\theta} \overset{\text{i.i.d.}}{\sim} f_k(x|\boldsymbol{\theta}) = \sum_{i=1}^{k} \lambda_i f(x|\xi_i)\,, \qquad j = 1, \ldots, n_x$$

where the component weights $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ are non-negative and sum to 1. The collection of the component-specific parameters is denoted by $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)$ and the collection of all parameters by $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\xi})$. Following a now standard representation (Frühwirth-Schnatter, 2001; Marin et al., 2005), each observation x_j from the sample can be assumed to originate from a specific unobserved component of f_k, denoted z_i, and the mixture inference problem can then be analysed as a missing data model, with discrete missing data $\mathbf{z} = (z_1, \ldots, z_{n_x})$, such that

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

*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

 DOI: 10.1214/15-BA970

The conditional distribution of $Z_j \in \{1, \ldots, k\}$ is then given by

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

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

In a Bayesian mixture modelling setup, the goal is to perform inference on the parameter $\boldsymbol{\theta}$, and the posterior distribution $\pi_k(\boldsymbol{\theta}|\mathbf{x})$ is usually approximated via MCMC methods. The likelihood function $p_k(\mathbf{x}|\boldsymbol{\theta})$ 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 moves very slowly if ever switches between the symmetric modes. Therefore, since the chain only explores a particular 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 removal 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)

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

where $\pi_k(\boldsymbol{\theta})$ is the selected prior for the k-component mixture. (In this paper, we assume that it is exchangeable with respect to its components.) Recall that the ratio of

evidences is a Bayes factor and is properly scaled to be compared with 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 said methods. Alternatively, estimating the number of components can proceed from a Bayesian non-parametric (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 $\pi_k(\boldsymbol{\theta}|\mathbf{x})$ (see the harmonic mean proposal of 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 in this paper 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 in this paper 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.

Our paper starts by 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. Experiments 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. It is derived from rewriting Bayes' theorem

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

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

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

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 a well-recognised fact that label switching is necessary for the above Rao–Blackwellised $\hat{\pi}_k(\boldsymbol{\theta}^o|\mathbf{x})$ to converge to the correct value. When $(\mathbf{z}^1, \ldots, \mathbf{z}^T)$ 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 $\mathrm{O}(\log k!)$, with no simple correction factor (Neal, 1999). Berkhof et al. (2003) suggested a generic correction by averaging $\widehat{\pi}_k(\boldsymbol{\theta}^o|\mathbf{x})$ over all possible permutations of the labels, hence forcing "perfect" label switching. The resulting approximation is expressed as

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

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

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

where the notational shortcut $\sigma(\boldsymbol{\theta}^o)$ means that the components of $\boldsymbol{\theta}^o$ are switched according to the permutation σ, i.e. $\sigma(\theta_1^o, \ldots, \theta_k^o) = (\theta_{\sigma(1)}^o, \ldots, \theta_{\sigma(k)}^o)$.

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 normalising 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 unnormalised posterior distribution $\pi_k^*(\boldsymbol{\theta}|\mathbf{x}) = \pi_k(\boldsymbol{\theta})p_k(\mathbf{x}|\boldsymbol{\theta})$, the bridge sampling identity is given by

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

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:

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

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

The convergence of bridge sampling (with the above optimal α) is trivial when $\pi_k^*(\boldsymbol{\theta}|\mathbf{x})$ and $q(\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$ 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 $(\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)})_{j=1}^{J_1}$, 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 $q(\boldsymbol{\theta}) = q(\boldsymbol{\lambda}, \boldsymbol{\xi})$ of the form

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

assuming $\{\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^{J_1}$ 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. A related MATLAB package, `bayesf`, by Frühwirth-Schnatter (2008) aggregates a bridge sampling estimator and an importance sampling estimator using the above q.

3 Novel importance sampling estimators

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

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

To provide a good approximation, the support of $q(\boldsymbol{\theta})$ must overlap with 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 number of permutations involved in $(\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)})_{j=1}^{J_1}$ (Frühwirth-Schnatter, 2004; Frühwirth-Schnatter, 2006).

In the following sections, the parameter $\boldsymbol{\theta}$ is decomposed into $(k+1)$ blocks $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \xi_1, \ldots, \xi_k)$. 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 Perrakis 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

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

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

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

This proposal is equivalent to generating $\boldsymbol{\theta}$ from $\pi_k(\boldsymbol{\theta}|\boldsymbol{\theta}^o, \mathbf{z}^o, \mathbf{x})$ and then operating a random permutation on the components of $\boldsymbol{\theta}$. The computational cost of producing

$\omega(\boldsymbol{\theta})$ in (5), hence $\widehat{\mathfrak{E}}(k)$, is thus multiplied by $k!$ under the provision that the support of (6) is sufficiently wide. If the tails of the density (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, indeed 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 $\pi_k(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}^o)$. This is indeed the practical version of simulating from an importance function with a support that is smaller than the support of the integrand and getting an erroneous approximation of the corresponding integral. In the current situation, since $\pi_k(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}^o)$ 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 $\{\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^J$ from the unconstrained posterior distribution. The new importance function is expressed as

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

Here, $\pi_k(\boldsymbol{\theta}|\sigma(\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)}), \mathbf{x})$ is a product of full conditional densities on each parameter in a Gibbs sampler representation and $\{\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^J$ 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.

We treat $(\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)})_{j=1}^J$ as parameters of q and denote them as $\{\varphi^{(j)}\}_{j=1}^J$. The density (7) then satisfies

$$q(\boldsymbol{\theta}) = \frac{1}{Jk!} \sum_{j=1}^{J} \sum_{i=1}^{k!} \pi_k(\boldsymbol{\theta}|\sigma_i(\boldsymbol{\varphi}^{(j)}), \mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{k!} \sum_{i=1}^{k!} h_{\sigma_i}(\boldsymbol{\theta}) \tag{8}$$

where $h_{\sigma_i}(\boldsymbol{\theta}) = \frac{1}{J} \sum_{j=1}^{J} \pi_k(\boldsymbol{\theta}|\sigma_i(\boldsymbol{\varphi}^{(j)}), \mathbf{x})$. Each of the densities $h_{\sigma_1}, \ldots, h_{\sigma_{k!}}$ has a support (i.e. a domain where it takes non-negligible values) denoted by $S_{\sigma_1}, \ldots, S_{\sigma_{k!}}$, re-

spectively, 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 $q(\boldsymbol{\theta})$ but not in generating a proposal $\boldsymbol{\theta}$ from q. For arbitrary permutation representations $\sigma_m, \sigma_c, \sigma_i \in \mathfrak{S}_k = \{\sigma_1, \ldots, \sigma_{k!}\}$ acting on both $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, the following holds for (7)

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

where $\sigma_m\sigma_c(\boldsymbol{\theta}) = \sigma_m(\sigma_c(\boldsymbol{\theta}))$. The full permutation representation set is invariant over an additional permutation representation σ_m (e.g. $\mathfrak{S}_k = \{\sigma_m\sigma_1, \ldots, \sigma_m\sigma_{k!}\}$) hence, $q(\sigma_c(\boldsymbol{\theta}))$ and $q(\sigma_m\sigma_c(\boldsymbol{\theta}))$ 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 $h_{\sigma_c}(\boldsymbol{\theta})$ 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 $h_\sigma(\boldsymbol{\theta})$'s need be evaluated to compute $q(\boldsymbol{\theta})$.

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 estimator (5) should therefore be more efficient. Both are designed to cover all of the $k!$ clusters, which are created by either symmetrising the labels of hyperparameter set $\{\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^{J}$ as in (8) or by randomly permuting the label of each $\{\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)}\}_{j=1}^{J_1}$ as in (4). Once $k!$ clusters of parameters of q 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

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

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 $(\kappa = 1, \ldots, k!)$

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

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 $q(\boldsymbol{\theta})$ as in (9) follows by a standard importance sampling identity

$$\mathfrak{E}(k) \quad = \quad \int_{S_q} \frac{\pi_k(\boldsymbol{\theta}) p_k(\mathbf{x}|\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\mathbf{x})} \left(\sum_{\kappa=1}^{k!} p(\kappa) q_\kappa(\boldsymbol{\theta}|\mathbf{x}) \right) \mathrm{d}\boldsymbol{\theta}$$

$$= \sum_{\kappa=1}^{k!} \int_{S_\kappa} \frac{\pi_k(\boldsymbol{\theta}) p_k(\mathbf{x}|\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\mathbf{x})} p(\kappa) q_\kappa(\boldsymbol{\theta}|\mathbf{x}) \mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{p(\boldsymbol{\theta},\kappa)}[\omega(\boldsymbol{\theta})] \tag{10}$$

leading to

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

where $\omega(\boldsymbol{\theta}) = \pi_k(\boldsymbol{\theta}) p_k(\mathbf{x}|\boldsymbol{\theta}) / q(\boldsymbol{\theta}|\mathbf{x})$, 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 $\boldsymbol{\theta}$ over S_κ for $\kappa = 1, \ldots, k!$ are equal. Given an arbitrary cluster, κ^o, the evidence is

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

Note that the corresponding estimator (Monte Carlo approximation based on T draws) for the above is exactly of the same form as 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 asymptotically converge to $\mathfrak{E}(k)$, and the Central Limit theorem also holds

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

where $V \overset{\text{def}}{=} V_1 = \mathrm{var}_{p(\boldsymbol{\theta},\kappa|\mathbf{x})}(\omega(\boldsymbol{\theta}))$ and $V \overset{\text{def}}{=} V_2 = \mathrm{var}_{q_{\kappa^o}(\boldsymbol{\theta}|\mathbf{x})}(\omega(\boldsymbol{\theta}))$ for (4) and (8), respectively. 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 importance functions approximate the joint posterior distribution. If $J_1 = Jk!$ and both importance functions provide a good approximation and $V_1 \approx V_2$ is expected.

4 Dual importance sampling using an 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 $\boldsymbol{\theta}$ and that proposals can be generated from a particular singular term $h_{\sigma_c}(\boldsymbol{\theta})$ of (8) without any loss of statistical efficiency. Given $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}) \sim h_{\sigma_c}(\boldsymbol{\theta})$, it is natural to consider whether or not all terms in $\{h_{\sigma_1}(\boldsymbol{\theta}^{(t)}), \ldots, h_{\sigma_{k!}}(\boldsymbol{\theta}^{(t)})\}$ 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 and our proposal.

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

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

If $\eta_{\sigma_i}(\boldsymbol{\theta})$ is close to zero, $h_{\sigma_i}(\boldsymbol{\theta})$ is negligible within $q(\boldsymbol{\theta})$, and on the opposite $\eta_{\sigma_i}(\boldsymbol{\theta}) \approx 1$ indicates a high contribution of $h_{\sigma_i}(\boldsymbol{\theta})$. In other words, if the supports of h_{σ_i} and h_{σ_c} do not overlap, $\eta_{\sigma_i} = 0$. As the support intersection gets larger, η_{σ_i} gets close to 1. The expected relative contribution of $h_{\sigma_i}(\boldsymbol{\theta})$

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

is estimated by

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

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

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

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

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

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 $\{\sigma_1, \ldots, \sigma_n\}$, n being defined as the smallest size that satisfies the condition $\widehat{\phi}_n < \tau$. Under this truncation, the computational efficiency obviously improves.

Note that $\mathfrak{A}(k)$ is determined under the assumption that most proposals ($\boldsymbol{\theta}^{(t)}$) are potentially generated from h_{σ_c} since the quality of an approximation is only guaranteed under this assumption. Due to the perfect symmetry of $q(\boldsymbol{\theta})$ over the $k!$ permutations, the choice of σ_c is obviously irrelevant to the computational efficiency.

If h_σ's are well separated, most of the $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_i}(\boldsymbol{\theta})]$'s are likely to be negligible and the size of $\mathfrak{A}(k)$ becomes small, i.e. substantial reduction in computing time ensues. One natural attempt towards this goal is through rearranging component labels of the terms $(\boldsymbol{\theta}^{(j)}, \mathbf{z}^{(j)})_{j=1}^{J}$ using a label switching removal technique with hopes that supports of the transformed $\pi_k(\cdot|\boldsymbol{\varphi}^{(j)}, \mathbf{x})$'s almost overlap. Therefore, the transformed h_σ's may be well separated (or may hardly overlap). Since the importance function (q) does not change through use of transformed h_σ's, the approximation to $\mathfrak{E}(k)$ remains valid and computing time may thus be reduced. The evidence estimate using such an approximation is detailed in the following algorithm:

Algorithm 1 Dual importance sampling algorithm with approximation

1 Randomly select $\{\mathbf{z}^{(j)}, \boldsymbol{\theta}^{(j)}\}_{j=1}^{J}$ from a Gibbs sequence and rearrange component labels using a label switching removal technique. Then construct $q(\boldsymbol{\theta})$ as in (8).

2 Derive $h_{\sigma_c}(\boldsymbol{\theta})$ and generate particles $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T} \sim h_{\sigma_c}(\boldsymbol{\theta})$.

3 Construct an approximation, $\tilde{q}(\boldsymbol{\theta})$, using the first M terms in $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$:

- **3.1** Compute $(h_{\sigma_1}(\boldsymbol{\theta}^{(t)}), \ldots, h_{\sigma_{k!}}(\boldsymbol{\theta}^{(t)}), \eta_{\sigma_1}(\boldsymbol{\theta}^{(t)}), \ldots, \eta_{\sigma_{k!}}(\boldsymbol{\theta}^{(t)}))$ for $t = 1, \ldots, M$ and $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}(\boldsymbol{\theta})], \ldots, \widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}(\boldsymbol{\theta})]$ as in (12).
- **3.2** Reorder the σ's so that $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}(\boldsymbol{\theta})] \geq \cdots \geq \widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}(\boldsymbol{\theta})]$.
- **3.3** Initialise $n = 1$ and compute $\tilde{q}_n(\boldsymbol{\theta}^{(1)}), \ldots, \tilde{q}_n(\boldsymbol{\theta}^{(M)})$ as in (13) and $\widehat{\phi}_n$ as in (14). If $\widehat{\phi}_{n=1} < \tau$, go to Step 4. Otherwise set $n = n + 1$ and update $\tilde{q}_n$ and $\widehat{\phi}_n$ until $\widehat{\phi}_n < \tau$.

4 Calculate $\tilde{q}_n(\boldsymbol{\theta}^{(M+1)}), \ldots, \tilde{q}_n(\boldsymbol{\theta}^{(T)})$ and replace $q(\boldsymbol{\theta}^{(1)}), \ldots, q(\boldsymbol{\theta}^{(T)})$ with $\tilde{q}_n(\boldsymbol{\theta}^{(1)}), \ldots, \tilde{q}_n(\boldsymbol{\theta}^{(T)})$ in (5) to estimate $\widehat{\mathfrak{E}}(k)$.

In Step 1, we followed the method suggested 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 $\tilde{q}_n(\boldsymbol{\theta})$. 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 therefore depend on $|\mathfrak{A}(k)|$, 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 performances of seven marginal likelihood estimators. The simulated datasets, D_1 and D_2, are:

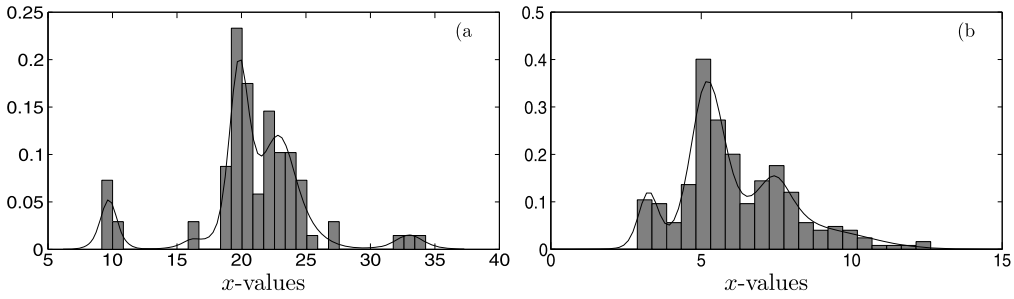

Figure 1: Histogram of the data against estimated six- and four-Gaussian mixture densities (solid line) for (a) the Galaxy dataset and (b) the fishery dataset, respectively.

(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 $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and standard deviation σ. 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 $q(\boldsymbol{\theta})$ with respect to M. Then we choose values for both M and τ. In Section 5.2, the performance of seven estimators are compared through a large simulation study, which confirms that the dual importance sampling is a reliable estimator with a lower demand in computation time.

5.1 Determining M and τ

The approximation set is constructed in two steps. First, we compute $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1(\boldsymbol{\theta})}]$, …, $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}(\boldsymbol{\theta})}]$, 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.

M	$\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_1}]$	$\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_2}]$	$\lvert\mathfrak{A}(k)\rvert$	$\widehat{\phi}_n$
10^2	1.0	1.89×10^{-102}	1	0
10^3	1.0	5.25×10^{-90}	1	0
10^4	1.0	4.62×10^{-91}	1	0
10^5	1.0	3.56×10^{-80}	1	0

Table 1: Estimates for $\{\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_i}]\}_{i=1}^{k!}$, $\lvert\mathfrak{A}(k)\rvert$ and $\widehat{\phi}_n$ against M for D_1 ($k = 2$). The prior for a variance parameter is $IG(2, 3)$. Note that $\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_1}] = 1$ due to rounding.

M	$\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_1}]$	$\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_2}]$	$\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_3}]$	$\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_4}]$	$\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_5}]$	$\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_6}]$	$\lvert\mathfrak{A}(k)\rvert$	$\widehat{\phi}_n$
10^2	1.0	3.56×10^{-16}	5.05×10^{-55}	4.64×10^{-65}	8.27×10^{-144}	9.53×10^{-160}	2	0
10^3	1.0	1.22×10^{-8}	3.01×10^{-49}	2.27×10^{-53}	3.08×10^{-125}	1.11×10^{-144}	2	0
10^4	1.0	2.03×10^{-8}	1.76×10^{-43}	4.87×10^{-49}	2.61×10^{-95}	8.31×10^{-136}	2	0
10^5	1.0	1.04×10^{-5}	2.27×10^{-39}	1.51×10^{-44}	4.30×10^{-87}	1.56×10^{-122}	2	0

Table 2: Estimates for $\{\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_i]\}_{i=1}^{k!}$, $\lvert\mathfrak{A}(k)\rvert$ and $\widehat{\phi}_n$ with respect to M for D_2 ($k = 3$). The prior for a variance parameter is $IG(2, 15)$. Note that $\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_1}] = 1$ due to rounding errors.

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 $(\widehat{\mathbb{E}}_{h_{\sigma_1}}[\eta_{\sigma_i}(\boldsymbol{\theta})])_{i=1}^{k!}$, and both $\lvert\mathfrak{A}(k)\rvert$ and $\widehat{\phi}$ are relatively stable against M. Under a natural lack of label switching, $q(\boldsymbol{\theta})$ seems to be well approximated using only $h_{\sigma_1}(\boldsymbol{\theta})$, 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, $h_{\sigma_1}(\boldsymbol{\theta})$ and $h_{\sigma_2}(\boldsymbol{\theta})$, significantly contribute to $q(\boldsymbol{\theta})$, as seen in Table 2. For the subsequent analyses in this paper, we chose $J = 10^2$, $M = 10^3$ and $\tau = 10^{-324}$.

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 lack of label switching;

$\widehat{\mathfrak{E}}_{Ch}$ Chib's method (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 $z_1^o, \ldots, z_n^o$ 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 $q(\boldsymbol{\theta})$;

$\widehat{\mathfrak{E}}_{DS}^{A}$ 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 and $J_1 = 100k!$ Gibbs samples in q;

$\widehat{\mathfrak{E}}_{BS}$ Bridge sampling (3), using $M_1 = M_2 = 5 \times 10^3$ samples and $q(\boldsymbol{\theta})$ as in (4) via 10 iterations. Label switching is imposed in hyperparameters $\{\boldsymbol{\theta}^{(j)}, z^{(j)}\}_{j=1}^{J_1}$ in q and $J_1 = 100k!$.

The marginal likelihood estimates (in log-scales) and the effective sample size (ESS) ratios, $R = \text{ESS}/T$, are summarised 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 (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 workload, except for $\widehat{\mathfrak{E}}_{DS}^{A}$.

Simulated mixture datasets

Mixture models of two and three components are fitted to D_1 and D_2, respectively. Regardless of the presence or absence 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 possibly smaller Monte Carlo variations as seen in Figures 2 and 3 and Table 4. When a suitable approximate for $q(\boldsymbol{\theta})$ is used for the dual importance sampling, no significant difference in the estimates and the effective sample sizes is observed. The mean sizes of $\mathfrak{A}(k)$ in Table 5 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 natural label switching ever present in Gibbs sequences), the number of evaluations in q is reduced almost by the maximal factor of $1/k!$. Computing time increases by a factor of $k!$ with k for most importance sampling based estimators in Table 6 and increases relatively slowly for Chib's methods. Overall, $\widehat{\mathfrak{E}}_{BS}$ requires the highest computing time.

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
$\widehat{\mathfrak{E}}_{DS}^{A}$	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 $\widehat{\mathfrak{E}}_{BS}$ includes 10 iterations.

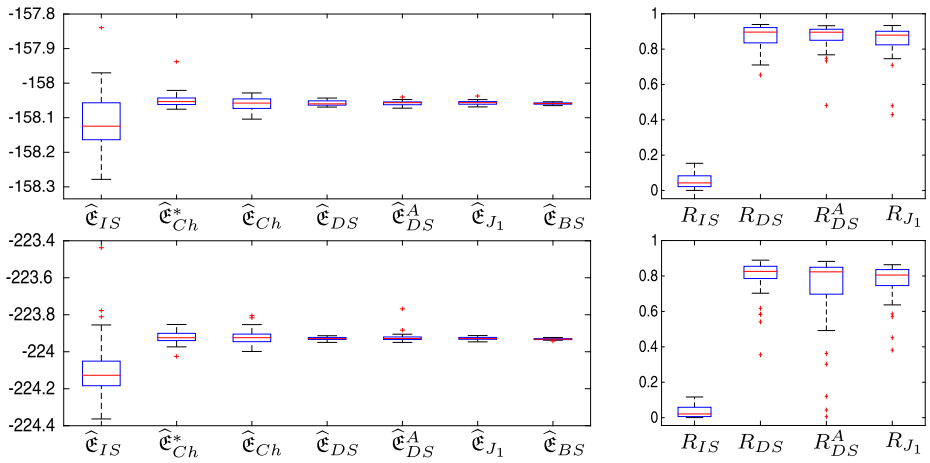

Figure 2: Boxplots of evidence estimates in log-scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with two and three Gaussian components are fitted to *(top)* D_1 and *(bottom)* D_2, respectively. The prior for $\{\sigma_i^2\}_{i=1}^k$ is $IG(2,3)$ and label switching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

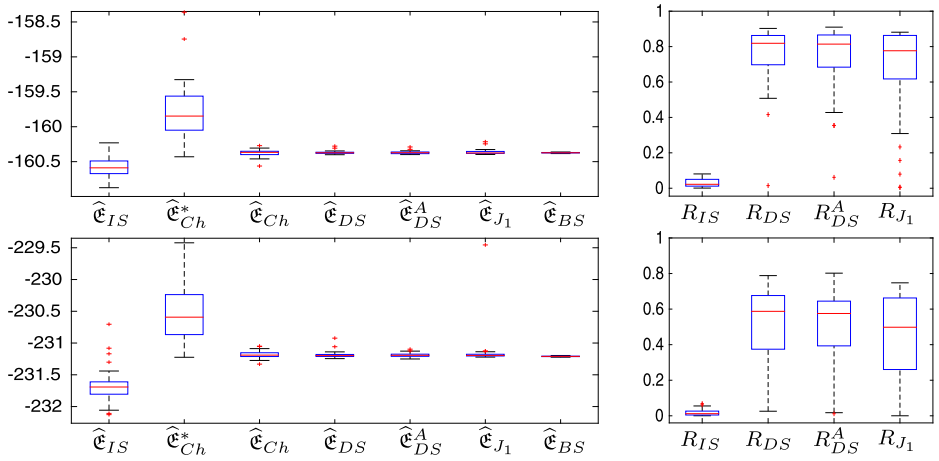

Figure 3: Boxplots of evidence estimates in log-scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with two and three Gaussian components are fitted to *(top)* D_1 and *(bottom)* D_2, respectively. The prior for $\{\sigma_i^2\}_{i=1}^k$ is $IG(2,15)$ and label switching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

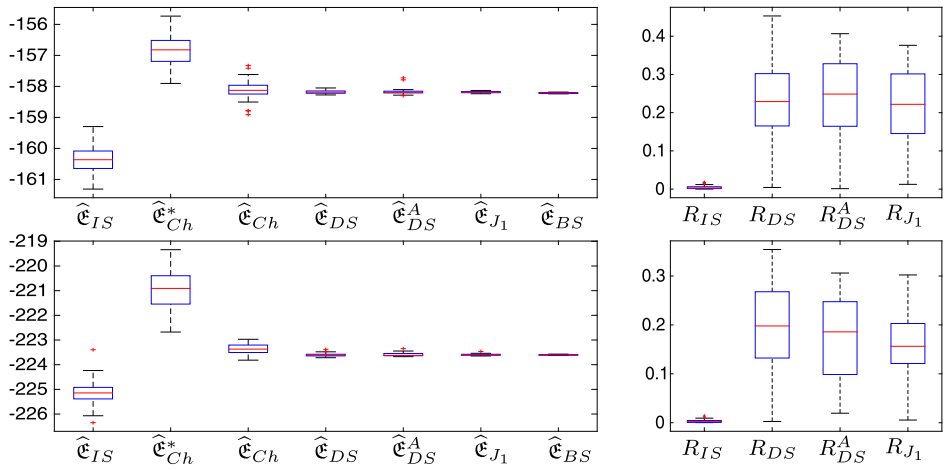

Figure 4: Boxplots of evidence estimates in log-scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with three and four Gaussian components are fitted to *(top)* D_1 and *(bottom)* D_2, respectively. The prior for $\{\sigma_i^2\}_{i=1}^k$ is $IG(2,3)$ and label switching occurs due to an extra component.

Data	$p(\sigma^2)$	k	$\hat{\mathfrak{E}}_{IS}$	$\hat{\mathfrak{E}}^{*}_{Ch}$	$\hat{\mathfrak{E}}_{Ch}$	$\hat{\mathfrak{E}}_{DS}$	$\hat{\mathfrak{E}}^{A}_{DS}$	$\hat{\mathfrak{E}}_{J_1}$	$\hat{\mathfrak{E}}_{BS}$
D_1	$IG(2,3)$	2	-158.0896	-158.0484	-158.0594	-158.0578	-158.0575	-158.0568	-158.0563
	$IG(2,15)$	2	-160.5695	-159.8051	-160.3762	-160.3689	-160.3700	-160.3618	-160.3721
	$IG(2,3)$	3	-160.3487	-156.8313	-158.1181	-158.1823	-158.1732	-158.1833	-158.2115
D_2	$IG(2,3)$	3	-224.0944	-223.9207	-223.9379	-223.9285	-223.9229	-223.9279	-223.9323
	$IG(2,15)$	3	-231.6750	-230.5445	-231.1845	-231.1876	-231.1902	-231.1539	-231.2115
	$IG(2,3)$	4	-225.1238	-220.9748	-223.3749	-223.6071	-223.5916	-223.5936	-223.6009

Table 4: Average values for 50 estimates for $\mathfrak{E}(k)$ per estimator.

When $\mathfrak{A}(k) < k!$, some reduction in the computing time of $\widehat{\mathfrak{E}}(k)^{A}_{DS}$ is observed and is due to ignoring zero function evaluations.

The above study considers the case when a mixture model is not overfitted, i.e. does not have superfluous components, and when label switching does not necessarily occur. In the event a mixture model is overfitted, label switching always occurs due to those extra (and unnecessary) components in a mixture representation. However, this does not mean that we cannot transform conditional densities associated with each of h_σ's to be closer, resorting to a label switching removal technique. For instance, mixture models with three and four components were fitted to the datasets D_1 and D_2, respectively. For those overfitted mixtures, performance phenomena similar to the above study are again observed in Figure 4 and Table 4; $\widehat{\mathfrak{E}}^{*}_{Ch}$ and $\widehat{\mathfrak{E}}_{IS}$ are quite off from the remaining estimators; an approximate set size $|\mathfrak{A}(k)|$ remains smaller than $k!$ in Table 5 and the corresponding computing time is reduced as shown by Table 6.

Disagreement in the values of $\widehat{\mathfrak{E}}_{IS}$ versus $\widehat{\mathfrak{E}}_{Ch}$ shows that an importance function may (unsurprisingly) fail to properly approximate $p_k(\mathbf{x}|\boldsymbol{\theta})\pi_k(\boldsymbol{\theta})$, resulting in an unreliable

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)

D	k	$k!$	$\lvert\mathfrak{A}_3(k)\rvert$	$\Delta(\mathfrak{A}_3)$
D_1	3	6	4.92 (1.26)	0.82 (0.21)
D_2	4	24	16.52 (8.28)	0.69 (0.34)

Table 5: 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 results using the priors $\sigma^2 \sim IG(2,3)$ and $\sigma^2 \sim IG(2,15)$, respectively. Estimates using overfitted models ($\sigma^2 \sim IG(2,3)$) are indicated by the subscript 3.

	D_1			D_2		
Estimator	$\triangle T_1$	$\triangle T_2$	$\triangle T_3$	$\triangle T_1$	$\triangle T_2$	$\triangle T_3$
$\widehat{\mathfrak{E}}^*_{Ch}$	0.64	0.91	1.01	1.19	1.22	1.50
$\widehat{\mathfrak{E}}_{Ch}$	0.64	0.70	0.99	1.09	1.09	1.56
$\widehat{\mathfrak{E}}_{IS}$	0.06	0.06	0.54	0.61	0.62	0.63
$\widehat{\mathfrak{E}}_{DS}$	0.52	0.51	1.56	1.73	1.67	7.06
$\widehat{\mathfrak{E}}^A_{DS}$	0.40	0.56	1.53	0.86	1.08	3.54
$\widehat{\mathfrak{E}}_{J_1}$	0.48	0.47	1.97	2.05	1.83	9.10
$\widehat{\mathfrak{E}}_{BS}$	0.87	0.84	2.49	2.42	2.38	10.75

Table 6: Elapsed time in seconds for evidence approximations in mixture models for D_1 and D_2 using a 2.5 GHz Intel Core i5 processor. Subscripts 1 and 2 of $\triangle T$ indicate results using the priors $\sigma^2 \sim IG(2,3)$ and $\sigma^2 \sim IG(2,15)$, respectively. Computing times for overfitted models ($\sigma^2 \sim IG(2,3)$) are indicated by the subscript 3.

estimate with large variation. Significantly small effective sample sizes (i.e. very small values for R_{IS}) back this observation. When label switching naturally occurs, as in the Gibbs sequence (either under the variance prior $IG(2,15)$ or due to extra components), $\widehat{\mathfrak{E}}^*_{Ch}$ disagrees with the other estimates, see Figures 3 and 4. Also 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).

Galaxy and fishery datasets

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

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

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 summarised in Fig-

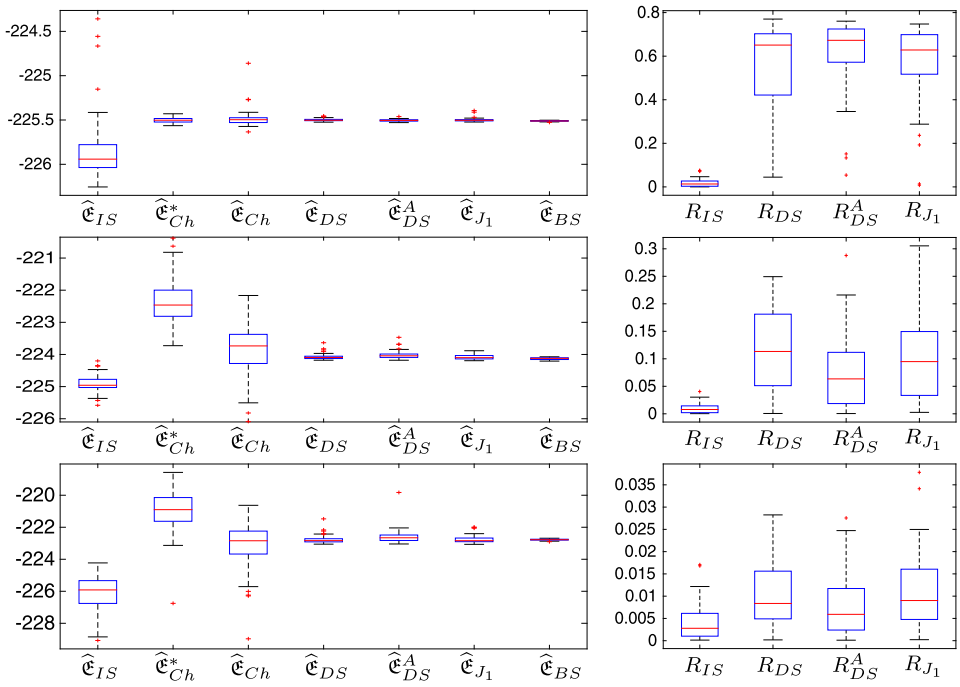

Figure 5: Simulation result for the galaxy dataset. Boxplots of evidence estimates in log-scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with *(top)* three, *(middle)* four, and *(bottom)* six Gaussian components are fitted. One outlier of $\widehat{\mathfrak{E}}_{Ch}$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

Data	k	$\hat{\mathfrak{E}}_{IS}$	$\hat{\mathfrak{E}}^{*}_{Ch}$	$\hat{\mathfrak{E}}_{Ch}$	$\hat{\mathfrak{E}}_{DS}$	$\hat{\mathfrak{E}}^{A}_{DS}$	$\hat{\mathfrak{E}}_{J_1}$	$\hat{\mathfrak{E}}_{BS}$
Fishery data	3	−519.6454	−519.3633	−519.5993	−519.3584	−519.3630	−519.4073	−519.3718
	4	−518.1560	−515.9706	−516.8680	−516.6662	−516.6196	−517.2511	−516.6378
Galaxy data	3	−225.8305	−225.5019	−225.4799	−225.4989	−225.5040	−225.5129	−225.4992
	4	−224.9167	−222.3848	−223.9332	−224.0716	−224.0109	−224.0754	−224.1287
	6	−226.1190	−221.0257	−223.1993	−222.7597	−222.5898	−222.7697	−222.7767

Table 7: Average values for 50 estimates for $\mathfrak{E}(k)$ per estimator.

ures 5 and 6 and Table 7. In general, a similar behaviour of $\log(\widehat{\mathfrak{E}}(k))$ and R is observed between the methods. For all cases, the dual importance sampling schemes ($\widehat{\mathfrak{E}}_{DS}$ and $\widehat{\mathfrak{E}}^{A}_{DS}$), $\widehat{\mathfrak{E}}_{J_1}$ and $\widehat{\mathfrak{E}}_{BS}$ agree with Chib's approach ($\widehat{\mathfrak{E}}_{Ch}$). Unless modes in 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. $\widehat{\mathfrak{E}}_{IS}$ is quite off from other estimates, due to a poor support for q.

Symptoms of the "curse of dimensionality" are observed. As k increases, the effective sample size decreases exponentially fast and the variation in the estimates increases. When $k = 6$, the variation in the estimates of $\widehat{\mathfrak{E}}_{Ch}$ is much larger than those based on importance sampling and the approximation ($\widehat{\mathfrak{E}}^{A}_{DS}$) with the current value for M is

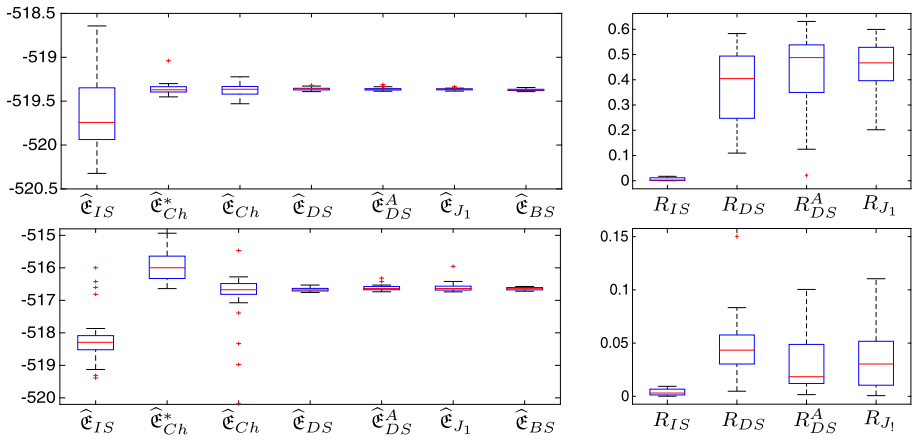

Figure 6: Simulation result for the fishery dataset. Boxplots of evidence estimates in log-scale *(left, middle)* and effective sample sizes ratios *(right)*. Mixture models with *(top)* three and *(bottom)* four Gaussian components are fitted. Two outliers of $\widehat{\mathfrak{E}}_{Ch}$ in the top-left panel are discarded.

slightly biased. Bridge sampling provides relatively stable estimates against k. Due to the exponential increase of $k!$, a fast increase in computing is observed for all estimators in Table 8.

The reduction in numbers of evaluated terms used to approximate $\widehat{\mathfrak{E}}(k)$ varies case by case, as shown in Table 9; the maximum reduction of 82% and the minimum of 57%. Particularly, a maximum computing time reduction of 59% compared to $\widehat{\mathfrak{E}}_{BS}$ is observed when $k = 4$ and $k = 6$ (see Table 8).

Estimator	Fishery data		Galaxy data		
	$k = 3$	$k = 4$	$k = 3$	$k = 4$	$k = 6$
$\widehat{\mathfrak{E}}^{*}_{Ch}$	1.16	1.54	1.06	1.68	2.13
$\widehat{\mathfrak{E}}_{Ch}$	1.18	1.61	1.05	1.88	7.19
$\widehat{\mathfrak{E}}_{IS}$	0.33	0.58	0.16	0.92	7.83
$\widehat{\mathfrak{E}}_{DS}$	1.91	8.47	1.74	8.61	396.81
$\widehat{\mathfrak{E}}^{A}_{DS}$	1.74	6.51	1.54	6.41	245.95
$\widehat{\mathfrak{E}}_{J_1}$	2.32	14.85	2.32	10.67	530.54
$\widehat{\mathfrak{E}}_{BS}$	2.82	13.79	2.76	12.53	610.18

Table 8: Elapsed time in seconds for evidences approximation of mixture models for fishery and galaxy datasets using a 2.5 GHz Intel Core i5 processor.

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	1.10 (0.30)	0.26 (0.04)
4	24	8.94 (4.56)	0.43 (0.17)
6	720	65.44 (27.38)	0.18 (0.03)

(b) Galaxy data

Table 9: Mean and standard deviation *(values in brackets)* of approximate set sizes, $\lvert\mathfrak{A}(k)\rvert$, and the reduction rate of a number of evaluated h-terms $\Delta(\mathfrak{A})$ as in (15) for (a) fishery and (b) galaxy datasets.

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 a posterior density. Importance sampling requires that the support of an importance function encompasses the support of a posterior density to perform properly. In the specific case of mixture models, missing some of modes in a posterior distribution is likely to produce an unsuitable support, hence a poor estimate of the evidence.

In our experiments, exchangeable priors are used and the posterior density exhibits $k!$ symmetrical terms. Two marginal likelihood estimators are proposed here and tested against other existing estimators. The first approach exploits the permutation of $\pi(\cdot|\mathbf{x}, \mathbf{z}^o)$ with a point-wise 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 evidence estimation for mixture models. Moreover, introducing a suitable and implementable approximation scheme, we show how to reduce the exponential increase in computational workload. The core idea of this approximation is to bypass negligible elements in the approximation thanks to the perfect symmetry of a 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 as well 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 characterise 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. MR0365969. 3

Ardia, D., Baştürk, N., Hoogerheide, L., and van Dijk, H. K. (2012). "A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood." *Computational Statistics and Data Analysis*, 56: 3398–3414. MR2943902. doi: http://dx.doi.org/10.1016/j.csda.2010.09.001. 3

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. MR1977735. 3, 4, 6

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

Celeux, G., Hurn, M., and Robert, C. P. (2000). "Computational and inferential difficulties with mixture posterior distributions." *Journal of the American Statistical Association*, 95(3): 957–979. MR1804450. doi: http://dx.doi.org/10.2307/2669477. 2

Chen, M.-H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer Series in Statistics, first edition. MR1742311. doi: http://dx.doi.org/10.1007/978-1-4612-1276-8. 5

Chib, S. (1995). "Marginal likelihoods from the Gibbs output." *Journal of the American Statistical Association*, 90: 1313–1321. MR1379473. 3, 4, 12

— (1996). "Calculating posterior distributions and modal estimates in Markov mixture models." *Journal of Econometrics*, 75: 79–97. MR1414504. doi: http://dx.doi.org/10.1016/0304-4076(95)01770-4. 3, 20

Chopin, N. (2002). "A sequential particle filter method for static models." *Biometrika*, 89(3): 539–552. MR1929161. doi: http://dx.doi.org/10.1093/biomet/89.3.539. 4

Chopin, N. and Robert, C. P. (2010). "Properties of nested sampling." *Biometrika*, 97: 741–755. MR2672495. doi: http://dx.doi.org/10.1093/biomet/asq021. 3, 4

Congdon, P. (2006). "Bayesian model choice based on Monte Carlo estimates of posterior model probabilities." *Computational Statistics and Data Analysis*, 50: 346–357. MR2201867. doi: http://dx.doi.org/10.1016/j.csda.2004.08.001. 3

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. MR1482122. doi: http://dx.doi.org/10.2307/2965554. 3

Diebolt, J. and Robert, C. (1994). "Estimation of finite mixture distributions through Bayesian sampling." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56: 363–375. MR1281940. 2

Doucet, A., Godsill, S., and Andrieu, C. (2000). "On sequential Monte Carlo sampling methods for Bayesian filtering." *Statistics and Computing*, 10: 197–208. 14

Escobar, M. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 3

Friel, N. and Pettitt, A. N. (2008). "Marginal likelihood estimation via power posteriors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70: 589–607. MR2420416. doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00650.x. 20

Friel, N. and Wyse, J. (2012). "Estimating the evidence: a review." *Statistica Neerlandica*, 66(3): 288–308. MR2955421. doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00515.x. 20

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. MR1952732. doi: http://dx.doi.org/10.1198/016214501750333063. 1, 2, 5

— (2004). "Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques." *Journal of Econometrics*, 7: 143–167. MR2076630. doi: http://dx.doi.org/10.1111/j.1368-423X.2004.00125.x. 5, 6

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics, first edition. MR2265601. 6, 12, 17

— (2008). *bayesf : Finite Mixture and Markov Switching Models*. MATLAB package version 2.0. http://statmath.wu.ac.at/ fruehwirth/monographie/book_matlab_version_2.0.pdf 6

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. MR1141740. 3, 4

Gelman, A. and Meng, X. L. (1998). "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling." *Statistical Science*, 13: 163–185. MR1647507. doi: http://dx.doi.org/10.1214/ss/1028905934. 3, 5

Geweke, J. (2012). "Interpretation and inference in mixture models: simple MCMC works." *Computational Statistics and Data Analysis*, 51: 3529–3550. MR2367818. doi: http://dx.doi.org/10.1016/j.csda.2006.11.026. 2

Green, P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 85(4): 711–732. MR1380810. doi: http://dx.doi.org/10.1093/biomet/82.4.711. 4

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. MR2182987. doi: http://dx.doi.org/10.1214/088342305000000016. 2, 11, 12

Jeffreys, H. (1939). *Theory of Probability*. Oxford, The Clarendon Press, first edition. 3

Marin, J. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, New York. MR2723361. 2

— (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. 3

— (2010b). "On resolving the Savage–Dickey paradox." *Electronic Journal of Statistics*, 4: 643–654. MR2660536. doi: http://dx.doi.org/10.1214/10-EJS564. 3

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. MR2490536. doi: http://dx.doi.org/10.1016/S0169-7161(05)25016-2. 1

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. 3, 6, 17

Meng, X. L. and Schilling, S. (2002). "Warp Bridge sampling." *Journal of Computational Graphical Statistics*, 11(3): 552–586. MR1938446. doi: http://dx.doi.org/10.1198/106186002457. 3, 5

Meng, X. L. and Wong, W. H. (1996). "Simulating ratios of normalizing constants via a simple identity." *Statistica Sinica*, 6: 831–860. MR1422406. 3, 5

Mira, A. and Nicholls, G. (2004). "Bridge estimation of the probability density at a point." *Statistica Sinica*, 14: 603–612. MR2059299. 5

Neal, R. M. (1999). "Erroneous results in Marginal likelihood from the Gibbs output." http://www.cs.toronto.edu/~radford/chib-letter.html 3, 4, 17

— (2001). "Annealed importance sampling." *Statistics and Computing*, 11: 125–139. MR1837132. doi: http://dx.doi.org/10.1023/A:1008923215028. 20

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 96(1): 3–48. MR1257793. 3

Papastamoulis, P. (2013). *label.switching: Relabelling MCMC outputs of mixture models*. R package version 1.2. http://CRAN.R-project.org/package=label.switching 2

Papastamoulis, P. and Iliopoulos, 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. MR2758306. doi: http://dx.doi.org/10.1198/jcgs.2010.09008. 2, 11

Papastamoulis, P. and Roberts, G. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95: 315–321. MR2409721. doi: http://dx.doi.org/10.1093/biomet/asm086. 2

Perrakis, 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. MR3210048. doi: http://dx.doi.org/10.1016/j.csda.2014.03.004. 6

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. MR2433201. 3

Rasmussen, C. E. (2000). "The Infinite Gaussian Mixture Model." In: *Advances in Neural Information Processing Systems 12*, 554–560. MIT Press. 3

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 (Statistical Methodology)*, 59(4): 731–792. MR1483213. doi: http://dx.doi.org/10.1111/1467-9868.00095. 2, 3, 4, 12, 17

Robert, C. and Marin, J.-M. (2008). "On some difficulties with a posterior probability approximation technique." *Bayesian Analysis*, 3(2): 427–442. MR2407433. doi: http://dx.doi.org/10.1214/08-BA316. 3

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, second edition. MR2080278. doi: http://dx.doi.org/10.1007/978-1-4757-4145-2. 2, 4

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. MR3173759. doi: http://dx.doi.org/10.1080/10618600.2012.735624. 2, 11

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. MR0898357. 3

— (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. 3

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. 3

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. MR1963393. doi: http://dx.doi.org/10.1198/016214502753479464. 3

Servidea, J. D. (2002). "Bridge sampling with dependent random draws: techniques and strategy." Ph.D. thesis, Department of Statistics, The University of Chicago. MR2717042. 5

Skilling, J. (2007). "Nested sampling for Bayesian computations." *Bayesian Analysis*, 1(4): 833–859. MR2282208. doi: http://dx.doi.org/10.1214/06-BA127. 4

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. MR1762903. doi: http://dx.doi.org/10.1214/aos/1016120364. 3

— (2000b). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62: 795–809. MR1796293. doi: http://dx.doi.org/10.1111/1467-9868.00265. 2, 12

Tierney, L. and Kadane, J. (1986). "Accurate approximations for posterior moments and marginal densities." *Journal of the American Statistical Association*, 81: 82–86. MR0830567. 3

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. MR1340514. 3

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. 5

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.