Brazilian Journal of Probability and Statistics
2019, Vol. 33, No. 4, 706–733
https://doi.org/10.1214/19-BJPS446

Keeping the balance—Bridge sampling for marginal likelihood estimation in finite mixture, mixture of experts and Markov mixture models

Sylvia Frühwirth-Schnatter

Vienna University of Economics and Business (WU), Austria

Abstract. Finite mixture models and their extensions to Markov mixture and mixture of experts models are very popular in analysing data of various kind. A challenge for these models is choosing the number of components based on marginal likelihoods. The present paper suggests two innovative, generic bridge sampling estimators of the marginal likelihood that are based on constructing balanced importance densities from the conditional densities arising during Gibbs sampling. The full permutation bridge sampling estimator is derived from considering all possible permutations of the mixture labels for a subset of these densities. For the double random permutation bridge sampling estimator, two levels of random permutations are applied, first to permute the labels of the MCMC draws and second to randomly permute the labels of the conditional densities arising during Gibbs sampling. Various applications show very good performance of these estimators in comparison to importance and to reciprocal importance sampling estimators derived from the same importance densities.

1 Introduction

Finite mixture models and their extensions to Markov mixture and mixture of experts models are very popular in analysing data of various kind. These models are useful for flexible modelling, density estimation and unsupervised clustering, see, for example, Frühwirth-Schnatter (2006) and Frühwirth-Schnatter, Celeux and Robert (2019) for a recent review. The various types of mixture models share a common structure insofar as it is supposed that N observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ are generated by K hidden groups/states. If the unknown group/state indicators $\mathbf{S} = (S_1, \ldots, S_N)$ are introduced as missing data, then the different model classes differ in their assumption concerning the distribution of the latent indicators $\mathbf{S}$.

For a finite mixture model, the indicators S_i are i.i.d. with $\Pr(S_i = k) = \eta_k$ and a mixture with weight distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ results as marginal distribution of $\mathbf{y}_i$. For a mixture of experts model, the indicators S_i are still independent, but the weight distribution $\Pr(S_i = k|\mathbf{x}_i)$ depends on covariates $\mathbf{x}_i$ and additional parameters $\boldsymbol{\gamma}$. For Markov mixture models, $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is a time series and

Key words and phrases. Markov chain Monte Carlo, model-based clustering, Gaussian mixtures, hierarchical priors, permutation sampling, importance sampling.

Received February 2019; accepted April 2019.

S_i is a hidden Markov chain with transition matrix $\boldsymbol{\xi}$. Given the group indicator S_i, for all three model classes $\mathbf{y}_i|S_i = k$ arises from a distribution $p(\mathbf{y}_i|\boldsymbol{\theta}_k)$ with group-specific parameter $\boldsymbol{\theta}_k$ that might also depend on covariates.

A challenge for any kind of mixture model is choosing the number K of hidden states/groups, see Celeux, Frühwirth-Schnatter and Robert (2019) for a comprehensive review. For finite mixtures, reversible jump MCMC methods (Richardson and Green, 1997) have been employed to sample from the posterior $p(K|\mathbf{y})$, however these methods are very challenging to implement. An attractive alternative to choose the number of hidden groups in a model-based clustering context are sparse finite mixture models (Malsiner Walli, Frühwirth-Schnatter and Grün, 2016). However, this approach does not allow comparisons across different model classes or different prior choices and so far has not been extended to hidden Markov and Markov switching models. A very general form of model selection can be achieved by comparing models and priors through marginal likelihoods and reliable estimators of the marginal likelihood are important for Bayesian model selection

Hence, in the present paper, we focus on Bayesian model choice among mixture models of increasing number of components K through the marginal likelihood $p(\mathbf{y}|K)$, defined as

$$p(\mathbf{y}|K) = \int p(\mathbf{y}|\boldsymbol{\vartheta}, K)\, p(\boldsymbol{\vartheta}|K)\, \mathrm{d}\boldsymbol{\vartheta}. \tag{1}$$

In (1), $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\omega})$ summarizes all unknown parameters, with $\boldsymbol{\omega}$ being a generic notation for the parameters in the weight distribution for all three types of mixture models considered in this paper. The marginal likelihood naturally penalises models with more mixture components (and more parameters), see, for example, Berger and Jefferys (1992); however, for mixture models it is not available in closed form and computational approximation methods become an integral part of model selection.

Frühwirth-Schnatter (2004) introduced simulation-based estimators such as importance sampling (Geweke, 1989), reciprocal importance sampling (Gelfand and Dey, 1994) or bridge sampling (Meng and Wong, 1996) to approximate the marginal likelihood for finite mixture and Markov switching models with moderate values of K. For such sampling-based techniques, one has to select for each K an importance density $q_K(\boldsymbol{\vartheta})$ which is easy to sample from and provides a rough approximation to the mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$. However, for mixture models, it is not at all straightforward to choose an appropriate importance density and the reliability of the resulting sampling-based estimators depends on several factors.

First, as shown by Frühwirth-Schnatter (2004), the tail behaviour of the importance density $q_K(\boldsymbol{\vartheta})$ compared to the mixture posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ matters. Whereas the (optimal) bridge sampling estimator (which will be reviewed in Section 2) is fairly robust in this respect, other sampling-based estimators are more

sensitive. For instance, importance sampling which is based on rewriting (1) as

$$p(\mathbf{y}|K) = \int \frac{p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K)}{q_K(\boldsymbol{\vartheta})} q_K(\boldsymbol{\vartheta})\,\mathrm{d}\boldsymbol{\vartheta}, \tag{2}$$

exhibits high standard errors, if $q_K(\boldsymbol{\vartheta})$ has thin tails compared to the mixture posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$.

Second, as pointed out by Lee and Robert (2016), the importance density $q_K(\boldsymbol{\vartheta})$ has to mimic the multimodality of the mixture posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ which is caused by the invariance of a mixture model with symmetric priors for the components to permutations of the mixture component labels, the so-called label switching problem. As proven in Rousseau, Grazian and Lee (2019), the number of symmetric modes in the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ tends to $K!$ as the number of observations N increases. A balanced importance density covers all modes of the posterior or, more formally, $q_K(\boldsymbol{\vartheta})$ is (nearly) invariant to permuting the labels of $\boldsymbol{\vartheta}$. If the importance density is unbalanced and several modes of the mixture posterior are not covered, then sampling-based estimators of the marginal likelihood are prone to be biased.

Several approaches have been suggested to ensure multimodality in the construction of the importance density also for increasing values of K. Frühwirth-Schnatter (2004) constructs the importance density from the output of random permutation posterior sampling (Frühwirth-Schnatter, 2001). However, as demonstrated in Celeux, Frühwirth-Schnatter and Robert (2019) for univariate Gaussian mixtures, the resulting bridge sampling estimator might be biased, despite its robustness to the tail behaviour. A first contribution of the present paper is to show that marginal likelihood estimators based on random permutation posterior sampling can be improved considerably by introducing a second level of random permutation during the construction of the importance density from the conditional densities arising during Gibbs sampling. This restores balance and yields the so-called double random permutation bridge sampling estimator.

An alternative approach is based on constructing perfectly balanced importance densities by considering all possible permutations of the labels, see, for example, Berkhof, van Mechelen and Gelman (2003) and Lee et al. (2009). Lee and Robert (2016) combine importance sampling with such a perfectly balanced importance density, calling the resulting estimator *dual importance sampling*. Celeux, Frühwirth-Schnatter and Robert (2019) show that a particularly stable estimator of the marginal likelihood, called full permutation bridge sampling estimator, is obtained for univariate Gaussian mixtures by combining (optimal) bridge sampling with a perfectly balanced importance density $q_K(\boldsymbol{\vartheta})$.

The main contribution of the present paper is to introduce such a full permutation bridge sampling estimator of the marginal likelihood for a much broader class of mixture models, including finite mixture models of many kinds, mixture of experts models as well as hidden Markov and Markov switching models. For each of

these model classes, we discuss in detail how to construct fully balanced importance densities. The various estimators are illustrated and compared for the various model classes for well-known data sets. We show that for all model classes considered very stable estimators of the marginal likelihood are obtained by combining (optimal) bridge sampling with a perfectly balanced importance density. On the other hand, dual importance sampling (Lee and Robert, 2016) exhibits larger standard errors than double random permutation and full permutation bridge sampling estimators in many cases, in particular for overfitting mixtures.

The rest of the paper is organized as follows. Section 2 reviews bridge sampling estimators and discusses the construction of the importance density from the outcome of Markov chain Monte Carlo sampling. To achieve balance in the importance density, Section 3 introduces double random and full permutation bridge sampling. The implementation of these estimators for finite mixtures, Markov mixtures and Markov switching models as well as mixture of experts models is outlined in Section 4 and illustrative applications are provided in Section 5. Section 6 concludes.

2 Bridge sampling approximations to the marginal likelihood

2.1 Bridge sampling estimators

Meng and Wong (1996) introduced a very general bridge sampling technique to estimate the marginal likelihood as the normalising constant of the non-normalized posterior $p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K)$, derived from Bayes' theorem. Let $q_K(\boldsymbol{\vartheta})$ be an approximation to the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ and let $\alpha(\boldsymbol{\vartheta})$ be a positive function such that $\int \alpha(\boldsymbol{\vartheta})q_K(\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\mathbf{y}, K)\,\mathrm{d}\boldsymbol{\vartheta} > 0$. Exploiting that

$$\int \alpha(\boldsymbol{\vartheta})q_K(\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\mathbf{y}, K)\,\mathrm{d}\boldsymbol{\vartheta} = \int \alpha(\boldsymbol{\vartheta})\frac{p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K)}{p(\mathbf{y}|K)}q_K(\boldsymbol{\vartheta})\,\mathrm{d}\boldsymbol{\vartheta},$$

yields the general bridge sampling estimator of the marginal likelihood:

$$p(\mathbf{y}|K) = \frac{\mathrm{E}_{q_K(\boldsymbol{\vartheta})}(\alpha(\boldsymbol{\vartheta})p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K))}{\mathrm{E}_{p(\boldsymbol{\vartheta}|\mathbf{y},K)}(\alpha(\boldsymbol{\vartheta})q_K(\boldsymbol{\vartheta}))},$$

provided that all expectations are well-defined.

Meng and Wong (1996) derived an optimal choice for $\alpha(\boldsymbol{\vartheta})$ which yields a bridge sampling estimator that requires i.i.d. draws $\boldsymbol{\vartheta}^{(l)}, l = 1, \ldots, L$ from the importance density $q_K(\boldsymbol{\vartheta})$ and i.i.d. draws from the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$. As Markov chain Monte Carlo (MCMC) draws $\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M$ from the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ are typically autocorrelated, Meng and Schilling (1996) defined an alternative optimal bridge sampling estimator $p_{\mathrm{BS}}(\mathbf{y}|K)$ based on following function $\alpha(\boldsymbol{\vartheta})$:

$$\alpha(\boldsymbol{\vartheta}) = 1/\big(L \cdot q_K(\boldsymbol{\vartheta}) + M_\star \cdot p(\boldsymbol{\vartheta}|\mathbf{y}, K)\big).$$

$M_\star$ is the effective sample size, estimated as $\hat{M}_\star = \min(M, M/\hat{\rho})$, where $\hat{\rho}$ is an estimator of the inefficiency factor of the posterior draws $f^{(m)} = p(\mathbf{y}|\boldsymbol{\vartheta}^{(m)}, K)p(\boldsymbol{\vartheta}^{(m)}|K)$. This definition of $\alpha(\boldsymbol{\vartheta})$ requires knowledge of the (unknown) normalizing constant $p(\mathbf{y}|K)$ to evaluate $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$. Using the estimator $\hat{p}_{\mathrm{IS}}(\mathbf{y}|K)$ (to be defined in (4)) as a starting value for $\hat{p}_{\mathrm{BS},0}(\mathbf{y}|K)$, the following recursion is applied until convergence to estimate the (optimal) bridge sampling estimator $\hat{p}_{\mathrm{BS}}(\mathbf{y}|K) = \lim_{t\to\infty} \hat{p}_{\mathrm{BS},t}(\mathbf{y}|K)$:

$$\hat{p}_{\mathrm{BS},t}(\mathbf{y}|K) = \frac{\frac{1}{L}\sum_{l=1}^{L} \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{(l)},K)p(\boldsymbol{\vartheta}^{(l)}|K)}{Lq_K(\boldsymbol{\vartheta}^{(l)})+\hat{M}_\star \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{(l)},K)p(\boldsymbol{\vartheta}^{(l)}|K)}{\hat{p}_{\mathrm{BS},t-1}(\mathbf{y}|K)}}}{\frac{1}{M}\sum_{m=1}^{M} \frac{q_K(\boldsymbol{\vartheta}^{(m)})}{Lq_K(\boldsymbol{\vartheta}^{(m)})+\hat{M}_\star \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{(m)},K)p(\boldsymbol{\vartheta}^{(m)}|K)}{\hat{p}_{\mathrm{BS},t-1}(\mathbf{y}|K)}}}. \tag{3}$$

Alternative estimators are obtained by other choices of $\alpha(\boldsymbol{\vartheta})$, for example, choosing $\alpha(\boldsymbol{\vartheta}) = 1/q_K(\boldsymbol{\vartheta})$ yields importance sampling as in (2). Based solely on the sample $\boldsymbol{\vartheta}^{(l)}, l = 1, \ldots, L$ from the importance density $q_K(\boldsymbol{\vartheta})$, the importance sampling estimator of the marginal likelihood is given by:

$$\hat{p}_{\mathrm{IS}}(\mathbf{y}|K) = \frac{1}{L}\sum_{l=1}^{L} \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{(l)}, K)p(\boldsymbol{\vartheta}^{(l)}|K)}{q_K(\boldsymbol{\vartheta}^{(l)})}. \tag{4}$$

Choosing, instead, $\alpha(\boldsymbol{\vartheta}) = 1/(p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K))$ yields the reciprocal importance sampling estimator (Gelfand and Dey, 1994):

$$p_{\mathrm{RI}}(\mathbf{y}|K) = \left(\mathrm{E}_{p(\boldsymbol{\vartheta}|\mathbf{y},K)}\left(\frac{q_K(\boldsymbol{\vartheta})}{p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K)}\right)\right)^{-1}.$$

This yields an estimator of the marginal likelihood solely based on the MCMC draws $\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M$ from the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$:

$$\hat{p}_{\mathrm{RI}}(\mathbf{y}|K) = \left(\frac{1}{M}\sum_{m=1}^{M} \frac{q_K(\boldsymbol{\vartheta}^{(m)})}{p(\mathbf{y}|\boldsymbol{\vartheta}^{(m)}, K)p(\boldsymbol{\vartheta}^{(m)}|K)}\right)^{-1}. \tag{5}$$

2.2 Defining importance densities for mixture analysis

Each of the estimators introduced in the previous section requires the choice of an importance density $q_K(\boldsymbol{\vartheta})$ for increasing K. As manual tuning of the importance density for each model under consideration is rather tedious, methods for choosing sensible importance densities in an unsupervised manner have been introduced. DiCiccio et al. (1997), for instance, suggested various methods to construct Gaussian importance densities from the MCMC output. However, the multimodality of the posterior density of a mixture model evidently forbids such a simple choice. Frühwirth-Schnatter (1995) is an early reference using Rao–Blackwellisation (Robert and Casella, 1999) to construct the importance density in an unsupervised manner from the MCMC output. She applied this idea to marginal

likelihood estimation for linear Gaussian state space models and extended this idea to finite mixture and Markov switching models in Frühwirth-Schnatter (2004).

For mixture models, a Rao–Blackwellised approximation of the posterior distribution of $\boldsymbol{\vartheta}$ based on introducing the latent allocations $\mathbf{S}$ as missing data yields:

$$p(\boldsymbol{\vartheta}|\mathbf{y}, K) = \int p(\boldsymbol{\vartheta}|\mathbf{S}, \mathbf{y}, K)\, p(\mathbf{S}|\mathbf{y}, K)\, \mathrm{d}\mathbf{S} \approx \frac{1}{M} \sum_{m=1}^{M} p\big(\boldsymbol{\vartheta}|\mathbf{S}^{(m)}, \mathbf{y}, K\big), \quad (6)$$

where $\mathbf{S}^{(m)}, m = 1, \ldots, M$ are M posterior draws of the latent allocations $\mathbf{S}$. The right-hand side of (6) is a mixture approximation of the posterior density $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ where the component densities $p(\boldsymbol{\vartheta}|\mathbf{S}^{(m)}, \mathbf{y}, K)$ arise in Gibbs sampling for mixture models (Diebolt and Robert, 1994), since $\boldsymbol{\vartheta}^{(m+1)}$ is drawn from $p(\boldsymbol{\vartheta}|\mathbf{S}^{(m)}, \mathbf{y}, K)$. If this conditional density arises from a well-known family of probability distributions, then its moments are available as a by-product of Gibbs sampling and can be stored easily, making the construction of an importance density based on the mixture approximation (6) fully automatic.

However, for mixture models there are several challenges with using (6) as importance density in bridge sampling techniques. First of all, the importance density $q_K(\boldsymbol{\vartheta})$ has to mimic the multimodality of the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ which results from invariance to label switching. Gibbs sampling might lead to (implicit) label switching in $\mathbf{S}^{(m)}$, meaning that the component densities $p(\boldsymbol{\vartheta}|\mathbf{S}^{(m)}, \mathbf{y}, K)$ in (6) will cover several posterior modes. However, even if M is very large, the resulting importance density $q_K(\boldsymbol{\vartheta})$ very likely is unbalanced, as (6) hardly ever covers *all* posterior modes equally well, if it is based on standard Gibbs sampling of $(\mathbf{S}, \boldsymbol{\vartheta})^{(m)}, m = 1, \ldots, M$. As noted earlier, balance of the importance density across all modes is important for obtaining reliable estimators for the marginal likelihood. Section 3 discusses various strategies to ensure that importance densities for mixture models are (nearly) balanced.

Second, despite introducing the latent states $\mathbf{S}$ as missing data, the conditional posterior $p(\boldsymbol{\vartheta}|\mathbf{S}, \mathbf{y}, K)$ is not available in closed form for many interesting mixture models. As will be shown in Section 4, a mixture approximation in the spirit of (6) can be constructed for these mixture models nevertheless, taking the form

$$q_K(\boldsymbol{\vartheta}) = \frac{1}{M} \sum_{m=1}^{M} q_K\big(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)}\big) q_K\big(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y}\big). \quad (7)$$

In (7), $\tilde{\mathbf{S}}^{(m)}$ is a generic notation summarizing all information needed to construct the mth component densities $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)})$ and $q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$ at the mth sweep of MCMC sampling. For instance, for non-Gaussian mixtures often a second level of data augmentation with latent variables $\mathbf{z}$ is introduced such that $p(\boldsymbol{\theta}_k|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$ with $\tilde{\mathbf{S}}^{(m)} = (\mathbf{S}^{(m)}, \mathbf{z}^{(m)})$ is of closed form. If $\tilde{\mathbf{S}}^{(m)} = \mathbf{S}^{(m)}$, then $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)}) = p(\boldsymbol{\omega}|\mathbf{S}^{(m)})$ and $q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y}) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\mathbf{S}^{(m)}, \mathbf{y})$ and (7) reduces to (6).

Provided balanced mixing across the posterior modes, (6) converges at a parametric speed toward the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ as M increases (Gelfand and Smith, 1990), whereas the density $q_K(\boldsymbol{\vartheta})$ defined in (7) remains an approximation to $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$, even if M goes to infinity, unless $\tilde{\mathbf{S}}^{(m)} = \mathbf{S}^{(m)}$. As choosing a large value of M makes the evaluation of $q_K(\boldsymbol{\vartheta})$ more expensive, an issue will be how to construct $q_K(\boldsymbol{\vartheta})$ from a subset of $Q < M$ component densities $q_K(\boldsymbol{\vartheta}|\tilde{\mathbf{S}}^{(q)}, \mathbf{y}) = q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(q)}) q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(q)}, \mathbf{y})$ in an efficient manner. On one hand, Q should be small for computational reasons, because $q_K(\boldsymbol{\vartheta})$ has to be evaluated for each of the Q components numerous times, for example, L times for the importance sampling estimator (4). On the other hand, to cover all symmetric modes of the posterior, a dramatically increasing value of Q proportional to $K!$ is required as K increases. Hence, estimators based on such an importance density are limited to fairly moderate values of K, say up to $K = 7$.

3 Achieving balance in the importance density

As discussed above, it is essential to construct the component densities $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)})$ and $q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$ in (7) from MCMC sampling such that $q_K(\boldsymbol{\vartheta})$ is nearly or even perfectly balanced. A perfectly balanced importance density $q_K(\boldsymbol{\vartheta})$ is entirely invariant to relabelling the components in $\boldsymbol{\vartheta}$. An efficient way to introduce multimodality in $q_K(\boldsymbol{\vartheta})$ and ensure (near) balance is to force label switching in a controlled manner.

3.1 Simple random and double random permutation estimators

An early suggestion to ensure multimodality in the construction of the importance density is based on random permutation posterior sampling (Frühwirth-Schnatter, 2004). A randomly selected permutation is applied at each sweep of MCMC sampling (Frühwirth-Schnatter, 2001) which creates explicit label switching in the component densities $p(\boldsymbol{\vartheta}|\mathbf{S}^{(m)}, \mathbf{y}, K)$ or, more generally, $q_K(\boldsymbol{\vartheta}|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$. A subset of these densities of size $Q < M$ is then used to construct the importance density $q_K^R(\boldsymbol{\vartheta})$ as a mixture approximation as in (7) and to compute the simple random permutation bridge sampling estimator $\hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$; see Algorithm 1 for details. Simple random permutation sampling has been applied in Frühwirth-Schnatter (2004) to finite mixture and Markov mixture models, and has been extended to mixtures of experts models in Frühwirth-Schnatter (2011).

Random permutation posterior sampling enhances mixing over all symmetric posterior modes and guarantees multimodality of the importance density $q_K^R(\boldsymbol{\vartheta})$ defined in (8). For regular cases, the number of modes in the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ tends to $K!$ as the number of observations N increases (Rousseau, Grazian and Lee, 2019). Choosing $Q = K!M_0$ ensures that on average each mode is visited M_0 times and M_0 components of the importance density are used to cover

Algorithm 1 Simple random permutation bridge sampling estimators

(a) Perform random permutation posterior sampling: for each $m = 1, \ldots, M$, conclude the mth sampling step by randomly drawing a permutation τ_m from $\mathcal{S}_K$, the set of the $K!$ permutations of the labels $\{1, \ldots, K\}$, and relabeling the mixture components: $(\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}, S_1^{(m)}, \ldots, S_N^{(m)})$ is substituted by $(\boldsymbol{\theta}_{\tau_m(1)}^{(m)}, \ldots, \boldsymbol{\theta}_{\tau_m(K)}^{(m)}, \tau_m^{-1}(S_1^{(m)}), \ldots, \tau_m^{-1}(S_N^{(m)}))$ and the parameters of the weight distribution are relabeled accordingly. For a finite mixture model, for instance, $(\eta_1^{(m)}, \ldots, \eta_K^{(m)})$ is substituted by $(\eta_{\tau_m(1)}^{(m)}, \ldots, \eta_{\tau_m(K)}^{(m)})$.

(b) Draw (without replacement) component densities $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(q)})q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(q)}, \mathbf{y})$ for $q = 1, \ldots, Q$ from the M component densities $q_K(\boldsymbol{\vartheta}|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$ derived from posterior sampling and construct following importance density:

$$q_K^R(\boldsymbol{\vartheta}) = \frac{1}{Q}\sum_{q=1}^{Q} q_K\big(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(q)}\big)q_K\big(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(q)}, \mathbf{y}\big). \tag{8}$$

(c) Use the importance density $q_K^R(\boldsymbol{\vartheta})$ to define the simple random permutation bridge sampling estimator $\hat{p}_{BS,R}(\mathbf{y}|K)$ from (3).

each posterior mode. Hence, for regular cases, $q_K^R(\boldsymbol{\vartheta})$ is (nearly) balanced for large enough values of M_0.

However, for less regular cases such as overfitting mixture models or small data sets, where more or less than $K!$ posterior modes are likely to be present, the importance density $q_K^R(\boldsymbol{\vartheta})$ tends to be unbalanced even for large values of Q. Whereas a perfectly balanced importance density is invariant to label switching (or a lack of it) in the MCMC draws $\boldsymbol{\vartheta}^{(m)}$, an imbalanced importance density can be quite sensitive in this respect. In addition, any lack of balance is amplified when K is large and Q approaches M, as the permutations underlying the MCMC draws are strongly tied to the permutations underlying the components densities. As a consequence, the MCMC draws and the components densities will over- or underrepresent the same modes. As recently shown in Celeux, Frühwirth-Schnatter and Robert (2019), this might create a bias in the corresponding bridge sampling estimator (3) for overfitting mixtures and larger values of K.

A surprisingly simple way to achieve (near) balance is introduced in Algorithm 2. It is based on drawing the Q components of the importance density $q_K^D(\boldsymbol{\vartheta})$ with replacement from the component densities arising during random permutation sampling and applying independent random permutations to each of these components. This so-called double random permutation bridge sampling estimator breaks the dependence between lack of balance in the posterior draws and lack of balance in the importance density, that can be observed for simple random permutation sampling.

Algorithm 2 Double random permutation bridge sampling estimators

(a) Perform random permutation posterior sampling as in Step (a) of Algorithm 1.
(b) Draw (with replacement) Q component densities $q_K(\boldsymbol{\vartheta}|\tilde{\mathbf{S}}^{(q)}, \mathbf{y})$ for $q = 1, \ldots, Q$ from the M component densities $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)})q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$ derived during posterior sampling for $m = 1, \ldots, M$.
(c) Draw (with replacement) a sequence of Q permutations $\rho_1, \ldots, \rho_Q$ from $\mathcal{S}_K$, the set of the $K!$ permutations of the labels $\{1, \ldots, K\}$, and construct following importance density:

$$q_K^D(\boldsymbol{\vartheta}) = \frac{1}{Q}\sum_{q=1}^{Q} q_K(\boldsymbol{\omega}|\rho_q(\tilde{\mathbf{S}}^{(q)}))q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\rho_q(\tilde{\mathbf{S}}^{(q)}), \mathbf{y}). \tag{9}$$

(d) Use the importance density $q_K^D(\boldsymbol{\vartheta})$ to define the double random permutation bridge sampling estimator $\hat{p}_{BS,D}(\mathbf{y}|K)$ from (3).

3.2 Full permutation estimators

As an alternative to random permutation sampling, several authors exploit full permutations to construct a completely balanced importance density, see, for example, Berkhof, van Mechelen and Gelman (2003), Frühwirth-Schnatter (2006) (Section 5.5.5) and Lee et al. (2009). The definition of such a fully symmetric importance density $q_K^F(\boldsymbol{\vartheta})$ is based on a mixture approximation as in (7). A small number M_0 of component densities $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(q)})q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(q)}, \mathbf{y}), q = 1, \ldots, M_0$, is selected from the M conditional densities $q_K(\boldsymbol{\vartheta}|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$, $m = 1, \ldots, M$, and expanded by including for each component s all $K!$ possible permutations.

This method yields the so-called full permutation bridge sampling estimator $\hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$, introduced in Algorithm 3. It should be noted that the importance density $q_K^F(\boldsymbol{\vartheta})$ is completely invariant to relabeling and therefore it is irrelevant whether the MCMC draws derived in Step (a) cover all posterior modes. Most notably, in (10) all symmetric modes are visited exactly M_0 times, leading to a symmetric, perfectly balanced importance density $q_K^F(\boldsymbol{\vartheta})$.

Note that both importance densities $q_K^F(\boldsymbol{\vartheta})$ and $q_K^D(\boldsymbol{\vartheta})$ can be used to define importance sampling estimators $\hat{p}_{\mathrm{IS},\bullet}(\mathbf{y}|K)$ as in (4) and reciprocal importance sampling estimators $\hat{p}_{\mathrm{RI},\bullet}(\mathbf{y}|K)$ as in (5). The dual importance sampling estimators of Lee and Robert (2016) results, if the importance density $q_K^F(\boldsymbol{\vartheta})$ is used in combination with (4) to define $\hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$.

The construction of $q_K^F(\boldsymbol{\vartheta})$ has in total $Q = M_0K!$ components, but is effectively based only on a small number M_0 of posterior draws $\tilde{\mathbf{S}}^{(s)}$. Hence, despite a possibly large number of terms Q in (10), the tail behaviour of the importance density is driven by the underlying M_0 components, meaning that $q_K^F(\boldsymbol{\vartheta})$ is only a rough approximation to the mixture posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ with possibly poor tail

Algorithm 3 Full permutation bridge sampling estimators

(a) Perform (standard) posterior sampling for $m = 1, \ldots, M$.

(b) Draw (with replacement) M_0 component densities $q_K(\boldsymbol{\vartheta}|\tilde{\mathbf{S}}^{(q)}, \mathbf{y})$ for $q = 1, \ldots, M_0$ from the M component densities $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)})q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$ derived during posterior sampling for $m = 1, \ldots, M$.

(c) For each $q = 1, \ldots, M_0$, define $K!$ expanded component densities by applying all possible permutations $\rho \in \mathcal{S}_K$:

$$q_K^F(\boldsymbol{\vartheta}) = \frac{1}{M_0} \sum_{q=1}^{M_0} \frac{1}{K!} \sum_{\rho \in \mathcal{S}_K} q_K\big(\boldsymbol{\omega}|\rho(\tilde{\mathbf{S}}^{(q)})\big) q_K\big(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\rho(\tilde{\mathbf{S}}^{(q)}), \mathbf{y}\big). \quad (10)$$

(d) Use the importance density $q_K^F(\boldsymbol{\vartheta})$ to define the full permutation bridge sampling estimator $\hat{p}_{BS,F}(\mathbf{y}|K)$ from (3).

behaviour for each single posterior mode. As a result, standard errors for dual importance sampling tend to be high due to their sensitivity to the tail behaviour of $q_K^F(\boldsymbol{\vartheta})$ in particular for overfitting models. As opposed to this, full permutation bridge sampling is very reliable also for overfitting mixtures, as it combines robustness with respect to the tail behaviour with robustness with respect to label switching.

Estimators based on full permutation bridge sampling have been applied to various specific model classes, including univariate Gaussian finite mixture models (Celeux, Frühwirth-Schnatter and Robert, 2019) as well as latent class models and finite Poisson mixture models (Frühwirth-Schnatter and Malsiner-Walli, 2019). We show in the present paper that full permutation bridge sampling is a very generic strategy and can be extended to more general finite mixtures (Section 4.1) and non-Gaussian mixtures (Section 4.3). Most importantly, full permutation bridge sampling can be extended in a natural way to hidden Markov and Markov switching models (Section 4.2) as well as mixture of experts models (Section 4.4).

4 Estimating marginal likelihoods in mixture analysis

4.1 Marginal likelihoods for finite mixtures

For finite mixtures, the prior often takes the form:

$$p(\boldsymbol{\vartheta}|K) = p(\boldsymbol{\eta}|K) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k),$$

where $p(\boldsymbol{\eta}|K) = \mathcal{D}_K(\boldsymbol{\eta}; e_0)$ is a symmetric Dirichlet distribution with hyperparameter e_0 and $p(\boldsymbol{\theta}_k)$ is conjugate to the conditional likelihood $p(\mathbf{y}|\boldsymbol{\theta}_k, \mathbf{S})$.[1] As a consequence, the complete-data posterior splits as $p(\boldsymbol{\vartheta}|\mathbf{S}, \mathbf{y}, K) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\mathbf{S}, \mathbf{y})\, p(\boldsymbol{\eta}|\mathbf{S})$. This implies that conditional on $\mathbf{S}$, the parameters defining the weight distribution $\boldsymbol{\eta}$ are independent from the mixture parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, when the importance density $q_K(\boldsymbol{\vartheta})$ is constructed using Rao-Blackwellisation as in (6). Very conveniently, this conditional independence given $\mathbf{S}^{(m)}$ is preserved, even if the complete-data posterior $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\mathbf{y}, \mathbf{S})$ is not of closed form. This justifies to construct $q_K(\boldsymbol{\vartheta})$ as a mixture approximation in the spirit of (7), using the conditionally independent components densities $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)})$ and $q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$.

The choice of $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)})$ depends on the model chosen for the indicators. For a finite mixture model, $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)}) = q_K(\boldsymbol{\eta}|\mathbf{S}^{(m)})$ is equal to the complete-data posterior $p(\boldsymbol{\eta}|\mathbf{S}^{(m)})$, taking the form of a Dirichlet distribution:

$$q_K\big(\boldsymbol{\eta}|\mathbf{S}^{(m)}\big) = \mathcal{D}\big(\boldsymbol{\eta}; e_1^{(m)}, \ldots, e_K^{(m)}\big), \tag{11}$$

where $e_k^{(m)} = e_0 + \sum_{i=1}^N I\{S_i^{(m)} = k\}$ with $I\{A\}$ being the indicator function for the event A.

The conditional independence yields a straightforward extension to Markov mixture and Markov switching models (Section 4.2) and can be extended to more general non-Gaussian mixtures (Section 4.3). Also for mixture of experts models conditional independence holds, however, no closed form posterior for the parameters $\boldsymbol{\omega}$ in the weight distribution exists. More details how to construct $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)})$ based on data augmentation are provided in Section 4.4.

Very conveniently, regardless of the specific type of mixture model, the construction of the component density $q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y})$ for the mixture parameters $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ follows the same strategy and only depends on the group specific density $p(\mathbf{y}_i|\boldsymbol{\theta}_k)$. The construction is straightforward for the one-block case, where the complete-data posterior $p(\boldsymbol{\theta}_k|\mathbf{S}^{(m)}, \mathbf{y})$ arises from a well-known distribution family. In this case, $\tilde{\mathbf{S}}^{(m)} = \mathbf{S}^{(m)}$ and

$$q_K\big(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y}\big) = \prod_{k=1}^K p\big(\boldsymbol{\theta}_k|\mathbf{S}^{(m)}, \mathbf{y}\big). \tag{12}$$

Consider, e.g. mixture analysis of count data, where the mixture components arise from a Poisson distribution, that is, $\mathbf{y}_i|S_i = k \sim \mathcal{P}(\mu_k)$. Based on the Gamma prior $\mu_k \sim \mathcal{G}(a_0, b_0)$, the full conditional posterior arises from the Gamma distribution $\mu_k|\mathbf{S}^{(m)}, \mathbf{y} \sim \mathcal{G}(a_k^{(m)}, b_k^{(m)})$, where

$$a_k^{(m)} = a_0 + \sum_{i=1}^N y_i I\big\{S_i^{(m)} = k\big\}, \qquad b_k^{(m)} = b_0 + \sum_{i=1}^N I\big\{S_i^{(m)} = k\big\}.$$

[1] More general hierarchical priors are discussed in Section 4.5.

Modifications are necessary when sampling from $p(\boldsymbol{\theta}_k|\mathbf{S},\mathbf{y})$ requires two (or even more) blocks, i.e. $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k,1},\ldots,\boldsymbol{\theta}_{k,B})$, and knowledge of $\mathbf{S}$ alone no longer leads to a simple closed-form density $p(\boldsymbol{\theta}_k|\mathbf{S}^{(m)},\mathbf{y})$. This is achieved by breaking the dependence between the various blocks $\boldsymbol{\theta}_{k,b}$ of $\boldsymbol{\theta}_k$ when constructing the components of $q_K(\boldsymbol{\vartheta})$. Frühwirth-Schnatter (1995) suggested to use the conditional densities in the transition kernel of the Gibbs sampler to construct $q_K(\boldsymbol{\theta}_{k,b}|\tilde{\mathbf{S}}^{(m)},\mathbf{y})$, where $\tilde{\mathbf{S}}^{(m)}$ includes $\mathbf{S}^{(m)}$ as well as the most recent values of all parameters appearing in the conditioning argument.

A typical example are multivariate Gaussian mixtures,

$$\mathbf{y}_i|S_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

under the non-conjugate prior $p(\boldsymbol{\theta}_k) = p(\boldsymbol{\mu}_k)p(\boldsymbol{\Sigma}_k)$ where $\boldsymbol{\theta}_k$ is sampled in two blocks from $p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k,\mathbf{S},\mathbf{y})$ and $p(\boldsymbol{\Sigma}_k|\boldsymbol{\mu}_k,\mathbf{S},\mathbf{y})$. Ignoring the dependence between $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, the component densities are constructed from conditionally independent densities,

$$q_K(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)},\mathbf{y}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{(m)},\mathbf{S}^{(m)},\mathbf{y})p(\boldsymbol{\Sigma}_k|\boldsymbol{\mu}_k^{(m-1)},\mathbf{S}^{(m)},\mathbf{y}),$$

given $\tilde{\mathbf{S}}^{(m)} = (\mathbf{S}^{(m)},\boldsymbol{\Sigma}_1^{(m)},\ldots,\boldsymbol{\Sigma}_K^{(m)},\boldsymbol{\mu}_1^{(m-1)},\ldots,\boldsymbol{\mu}_K^{(m-1)})$.

Estimators of the marginal likelihood based on double random permutation sampling (Algorithm 2) as well as full permutation sampling (Algorithm 3) are easily implemented. Given a permutation $\rho = (\rho(1),\ldots,\rho(K))$, the labels of the component densities $q_K(\boldsymbol{\eta}|\mathbf{S})$ and $q_K(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K|\tilde{\mathbf{S}},\mathbf{y})$ are permuted by reordering the labels of the corresponding complete-data moments according to ρ. For a finite mixture model, $q_K(\boldsymbol{\eta}|\rho(\mathbf{S}^{(m)}))$ is simply obtained by permuting the labels of the Dirichlet distribution (11):

$$q(\boldsymbol{\eta}|\rho(\mathbf{S}^{(m)})) = \mathcal{D}(\boldsymbol{\eta};e_{\rho(1)}^{(m)},\ldots,e_{\rho(K)}^{(m)}).$$

The mixture parameter component densities $q_K(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K|\rho(\tilde{\mathbf{S}}^{(m)}),\mathbf{y})$ are easily obtained by permuting the moments of the complete-data densities of $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K$. The precise details, however, depend on the specific mixture distribution. For mixtures of Poisson distributions, for instance, where $\mu_k|\mathbf{S}^{(m)},\mathbf{y} \sim \mathcal{G}(a_k^{(m)},b_k^{(m)})$, we simply obtain:

$$q_K(\mu_1,\ldots,\mu_K|\rho(\tilde{\mathbf{S}}^{(m)}),\mathbf{y}) = \prod_{k=1}^{K} \mathcal{G}(\mu_k;a_{\rho(k)}^{(m)},b_{\rho(k)}^{(m)}).$$

4.2 Marginal likelihoods for hidden Markov and Markov switching models

Estimators of the marginal likelihood based on double random permutation sampling (Algorithm 2) as well as full permutation sampling (Algorithm 3) are introduced for this model class in the present paper and provide a considerable improvement over simple random permutation sampling estimators as in Algorithm 1 (Frühwirth-Schnatter, 2004). Both estimators are easily implemented.

The construction of $q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K | \tilde{\mathbf{S}}^{(m)}, \mathbf{y})$ follows exactly Section 4.1, whereas the component density for the weight distribution is substituted by a component density $q_K(\boldsymbol{\xi}|\mathbf{S})$ for the transition matrix $\boldsymbol{\xi}$ of the hidden Markov chain. The prior $p(\boldsymbol{\xi})$ is defined row wise as $\boldsymbol{\xi}_{k,\cdot} \sim \mathcal{D}(e_{k1}^0, \ldots, e_{kK}^0)$ where $e_{kk}^0 \equiv e_p$ for all k and $e_{kj}^0 \equiv e_t$ for all $k \neq j$ to ensure invariance with respect to relabelling the states of S_i. The initial value S_0 of the hidden Markov chain is often assumed to arise from the ergodic distribution $\boldsymbol{\eta}_{\boldsymbol{\xi}}$ corresponding to the transition matrix $\boldsymbol{\xi}$. The complete-data posterior $p(\boldsymbol{\xi}|\mathbf{S}^{(m)})$ is given by:

$$p(\boldsymbol{\xi}|\mathbf{S}^{(m)}) = p(S_0^{(m)}|\boldsymbol{\eta}_{\boldsymbol{\xi}}) \prod_{k=1}^{K} p(\boldsymbol{\xi}_{k,\cdot}|\mathbf{S}^{(m)}),$$

where $p(\boldsymbol{\xi}_{k,\cdot}|\mathbf{S}^{(m)}) = \mathcal{D}(\boldsymbol{\xi}_{k,\cdot}; e_{k1}^{(m)}, \ldots, e_{kK}^{(m)})$ is equal to a Dirichlet distribution with

$$e_{kj}^{(m)} = e_{kj}^0 + N_{kj}^{(m)}, \qquad N_{kj}^{(m)} = \sum_{i=1}^{N} I\{S_{i-1}^{(m)} = k, S_i^{(m)} = j\}.$$

For simplicity, construction of the component density $q_K(\boldsymbol{\xi}|\mathbf{S}^{(m)})$ is based on ignoring the information in the prior $p(S_0^{(m)}|\boldsymbol{\eta}_{\boldsymbol{\xi}})$:

$$q_K(\boldsymbol{\xi}|\mathbf{S}^{(m)}) = \prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\xi}_{k,\cdot}; e_{k1}^{(m)}, \ldots, e_{kK}^{(m)}). \tag{13}$$

If S_0 is independent of $\boldsymbol{\xi}$, i.e. $p(S_0|\boldsymbol{\xi}) = p(S_0)$, then $q_K(\boldsymbol{\xi}|\mathbf{S}^{(m)})$ is identical with the complete-data posterior $p(\boldsymbol{\xi}|\mathbf{S}^{(m)})$.

Given a permutation $\rho = (\rho(1), \ldots, \rho(K))$, $q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\rho(\tilde{\mathbf{S}}^{(m)}), \mathbf{y})$ is permuted as in Section 4.1, whereas $q_K(\boldsymbol{\xi}|\rho(\mathbf{S}^{(m)}))$ is obtained by permuting the rows and the labels of the Dirichlet distribution (13) in the following way:

$$q_K(\boldsymbol{\xi}|\rho(\mathbf{S}^{(m)})) = \prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\xi}_{k,\cdot}; e_{\rho(k),\rho(1)}^{(m)}, \ldots, e_{\rho(k),\rho(K)}^{(m)}).$$

4.3 Marginal likelihoods for non-Gaussian mixtures

The methods discussed so far can be extended to non-Gaussian mixture models, where the complete-data likelihood $p(\boldsymbol{\theta}_k|\mathbf{S}, \mathbf{y})$ does not arise from a well-known distribution family. Examples include mixtures of skew-normal distributions and mixtures of generalized linear models. Data augmentation introducing (auxiliary) latent variables $\mathbf{z}$, in addition to $\mathbf{S}$, often leads to a Gibbs sampling scheme, where the complete-data posterior $p(\boldsymbol{\theta}_k|\mathbf{S}, \mathbf{z}, \mathbf{y})$ arises from a well-known distribution family. This allows to construct importance densities through Rao-Blackwellisation as in the previous sections also for non-Gaussian mixtures by

conditioning on $\tilde{\mathbf{S}} = (\mathbf{S}, \mathbf{z})$:

$$q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K | \tilde{\mathbf{S}}^{(m)}, \mathbf{y}) = \prod_{k=1}^{K} p(\boldsymbol{\theta}_k | \mathbf{S}^{(m)}, \mathbf{z}^{(m)}, \mathbf{y}). \tag{14}$$

Note that the sampling-based estimators of the marginal likelihood introduced in Section 2 are still based on the mixture likelihood $p(\mathbf{y}|\boldsymbol{\vartheta})$ as before, without conditioning on $\mathbf{z}$ or $\mathbf{S}$.

Using such an importance density, marginal likelihoods were approximated through simple random permutation bridge sampling estimators as in Algorithm 1 for mixtures of GLMs based on the Poisson and the negative binomial distribution (Frühwirth-Schnatter et al., 2009) and for univariate skew-normal and skew-t mixtures (Frühwirth-Schnatter and Pyne, 2010). In the present paper, estimators of the marginal likelihood based on double random permutation sampling (Algorithm 2) as well as full permutation sampling (Algorithm 3) are introduced as an interesting improvement.

Consider, for instance, a mixture of generalized linear models (GLMs), where the component densities $p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}_k)$ depend on covariates $\mathbf{x}_i$ through mixture regression parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$. Data augmentation through latent variables $\mathbf{z}$ together with a Gaussian prior for $\boldsymbol{\beta}_k$ leads to conditionally Gaussian posteriors $\boldsymbol{\beta}_k|\mathbf{S}^{(m)}, \mathbf{z}^{(m)}, \mathbf{y} \sim \mathcal{N}(\mathbf{b}_k^{(m)}, \mathbf{B}_k^{(m)})$. Such data augmentation methods include auxiliary mixture sampling (Frühwirth-Schnatter et al., 2009) and Polya-Gamma sampling (Polson, Scott and Windle, 2013). The posterior draws $\tilde{\mathbf{S}}^{(m)} = (\mathbf{S}^{(m)}, \mathbf{z}^{(m)})$ can be used to define component densities for the mixture regression parameters:

$$q_K(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K | \tilde{\mathbf{S}}^{(m)}, \mathbf{y}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\beta}_k; \mathbf{b}_k^{(m)}, \mathbf{B}_k^{(m)}). \tag{15}$$

Given a permutation $\rho = (\rho(1), \ldots, \rho(K))$ a permuted component simply reads

$$q_K(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K | \rho(\tilde{\mathbf{S}}^{(m)}), \mathbf{y}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\beta}_k; \mathbf{b}_{\rho(k)}^{(m)}, \mathbf{B}_{\rho(k)}^{(m)}).$$

4.4 Marginal likelihoods for mixture of experts models

The weight distribution $(\eta_1(\mathbf{x}_i), \ldots, \eta_K(\mathbf{x}_i))$ of a mixture of experts (ME) model depends for each observation $\mathbf{y}_i$ on covariates and is typically given by a multinomial logit (MNL) model:

$$\Pr(S_i = k|\mathbf{x}_i) = \eta_k(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i \boldsymbol{\gamma}_k)}{\sum_{k'=1}^{K} \exp(\mathbf{x}_i \boldsymbol{\gamma}_{k'})}, \tag{16}$$

where $\mathbf{x}_i$ is a row vector containing the covariates (including a constant) and $\boldsymbol{\gamma}_k$ are unknown regression parameters. One category, for example, $k_0 = 1$, is considered

as baseline with $\boldsymbol{\gamma}_{k_0} = 0$ for identifiability reasons; see Gormley and Frühwirth-Schnatter (2019) for a recent review of ME models. No sparse finite mixture framework has been developed for mixture of experts models sofar and only a few papers discuss marginal likelihood estimation. In the present paper, we introduce marginal likelihood estimators for ME models based on double random and full permutation sampling.

The MNL model (16) is a further example of a non-Gaussian model, where data augmentation based on latent variables $\mathbf{z}$ yields conditionally Gaussian posteriors $\boldsymbol{\gamma}_k|\boldsymbol{\gamma}_{-k}, \mathbf{S}, \mathbf{z} \sim \mathcal{N}(\mathbf{a}_k, \mathbf{A}_k)$ for simple MCMC updates of the weight parameters $(\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K)$ in a ME model. Data augmentation methods such as auxiliary mixture sampling (Frühwirth-Schnatter and Frühwirth, 2010) and Polya-Gamma sampling (Polson, Scott and Windle, 2013) allow to construct an importance density $q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}})$ for the weight parameters $\boldsymbol{\omega} = \{\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K\}$ in a similar manner as in Section 4.3, based on further blocking and assuming independence across blocks. Conditional on $\tilde{\mathbf{S}}^{(m)} = (\boldsymbol{\gamma}_{-k}^{(m)}, \mathbf{S}^{(m)}, \mathbf{z}^{(m)})$, where $\boldsymbol{\gamma}_{-k}^{(m)} = (\boldsymbol{\gamma}_{<k}^{(m)}, \boldsymbol{\gamma}_{>k}^{(m-1)})$, we obtain:

$$q_K(\boldsymbol{\omega}|\tilde{\mathbf{S}}^{(m)}) = \prod_{k=2}^{K} p(\boldsymbol{\gamma}_k|\boldsymbol{\gamma}_{-k}^{(m)}, \mathbf{z}^{(m)}, \mathbf{S}^{(m)}) = \prod_{k=2}^{K} \mathcal{N}(\boldsymbol{\gamma}_k; \mathbf{a}_k^{(m)}, \mathbf{A}_k^{(m)}). \tag{17}$$

Based on (17), Frühwirth-Schnatter (2011) used simple random permutation sampling estimators as in Algorithm 1 to compute marginal likelihoods for ME models. Alternatively, a fully balanced importance density $q_K^F(\boldsymbol{\vartheta})$ can be constructed as in Algorithm 3:

$$q_K^F(\boldsymbol{\vartheta}) = \frac{1}{M_0} \sum_{q=1}^{M_0} \frac{1}{K!} \sum_{\rho \in \mathcal{S}_K} \prod_{k=2}^{K} q_K(\boldsymbol{\gamma}_k|\rho(\tilde{\mathbf{S}}^{(q)})) q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\rho(\tilde{\mathbf{S}}^{(q)}), \mathbf{y}), \tag{18}$$

where the construction of $q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(q)}, \mathbf{y})$ follows exactly Section 4.1.

As noted by Frühwirth-Schnatter et al. (2012), special attention has to be paid to the correct relabelling of the coefficients $\boldsymbol{\gamma}_k$ in the MNL model (16) when applying a permutation ρ. This affects both permuting the labels during MCMC sampling in Step (a) of Algorithms 1 and 2 and constructing the importance density by permuting the components densities in Step (c) of Algorithms 2 and 3.

To relabel the weight distribution of an ME model for a given permutation ρ, define $\eta_k^\star(\mathbf{x}_i) = \eta_{\rho(k)}(\mathbf{x}_i)$ for $k = 1, \ldots, K$. The coefficients $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$ and $(\boldsymbol{\gamma}_1^\star, \ldots, \boldsymbol{\gamma}_K^\star)$ defining, respectively, the MNL models $\eta_k(\mathbf{x}_i)$ and $\eta_k^\star(\mathbf{x}_i)$ are related through

$$\mathbf{x}_i \boldsymbol{\gamma}_k^\star = \log\left[\frac{\eta_k^\star(\mathbf{x}_i)}{\eta_{k_0}^\star(\mathbf{x}_i)}\right] = \log\left[\frac{\eta_{\rho(k)}(\mathbf{x}_i)}{\eta_{\rho(k_0)}(\mathbf{x}_i)}\right] = \mathbf{x}_i(\boldsymbol{\gamma}_{\rho(k)} - \boldsymbol{\gamma}_{\rho(k_0)}).$$

Given ρ, the coefficients are permuted in the following way:

$$\boldsymbol{\gamma}_k^\star = \boldsymbol{\gamma}_{\rho(k)} - \boldsymbol{\gamma}_{\rho(k_0)}, \quad k = 1, \ldots, K. \tag{19}$$

This implies that $\boldsymbol{\gamma}_{k_0}^{\star} = 0$ and ensures that the baseline k_0 remains the same, despite relabelling. For $K = 2$, the signs of all coefficients of $\boldsymbol{\gamma}_2$ are simply flipped if $\rho = (2, 1)$, and remain unchanged otherwise.

Random permutation posterior sampling applies such relabelling at each sweep m during MCMC sampling using $\rho = \tau_m$. Correct relabelling of the densities $q_K(\boldsymbol{\gamma}_k|\rho(\tilde{\mathbf{S}}^{(q)}))$ in (18) proceeds as follows. For $k = k_0$, $q_K(\boldsymbol{\gamma}_{k_0}|\rho(\tilde{\mathbf{S}}^{(q)}))$ degenerates to a point mass at 0, as expected. For $k \neq k_0$, due to (19), $q_K(\boldsymbol{\gamma}_k|\rho(\tilde{\mathbf{S}}^{(q)}))$ is Gaussian with following moments:

$$q_K(\boldsymbol{\gamma}_k|\rho(\tilde{\mathbf{S}}^{(q)})) = \mathcal{N}(\boldsymbol{\gamma}_k; \mathbf{a}_{\rho(k)}^{(q)} - \mathbf{a}_{\rho(k_0)}^{(q)}, \mathbf{A}_{\rho(k)}^{(q)} + \mathbf{A}_{\rho(k_0)}^{(q)}).$$

4.5 Marginal likelihoods under hierarchical priors

For all kind of mixture models, the prior for the group-specific parameters often takes the following hierarchical form:

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|K) = p(\boldsymbol{\psi}) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k|\boldsymbol{\psi}). \tag{20}$$

For the one-block case, $p(\boldsymbol{\theta}_k|\boldsymbol{\psi})$ is conditionally conjugate to the conditional likelihood $p(\mathbf{y}|\boldsymbol{\theta}_k, \mathbf{S})$. Further blocking is needed, if the conditional priors within each block enjoy this property. For random hyperparameters $\boldsymbol{\psi}$, a hierarchical prior $p(\boldsymbol{\psi})$ is employed and posterior sampling is based on adding a block for sampling $\boldsymbol{\psi}^{(m)}$ from $p(\boldsymbol{\psi}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. Integrating over $p(\boldsymbol{\psi})$ yields a joint marginal prior $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|K)$ which usually has a closed form.

Also for hierarchical priors, marginal likelihood estimation is based on (1) and operates in the marginal space where $\boldsymbol{\psi}$ is integrated out. For the various bridge sampling estimators, the prior marginal $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|K)$ has to be evaluated at all draws $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ from the posterior or the importance density. This can be done using the candidate's formula, see, for example, Chib (1995):

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|K) = \frac{p(\boldsymbol{\psi}^{\star}) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k|\boldsymbol{\psi}^{\star})}{p(\boldsymbol{\psi}^{\star}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)}, \tag{21}$$

where $\boldsymbol{\psi}^{\star}$ is an arbitrary parameter value, for example, a draw from $p(\boldsymbol{\psi}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$.

While the various bridge sampling estimators operate in the marginal space where $\boldsymbol{\psi}$ is integrated out, the components of the importance density are constructed conditional on $\boldsymbol{\psi}$ to keep sampling from $q_K(\boldsymbol{\vartheta})$ simple. For instance, in the one-block case, $\tilde{\mathbf{S}}^{(m)} = (\mathbf{S}^{(m)}, \boldsymbol{\psi}^{(m)})$ is used, yielding

$$q_K(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\tilde{\mathbf{S}}^{(m)}, \mathbf{y}) = \prod_{k=1}^{K} q_K(\boldsymbol{\theta}_k|\tilde{\mathbf{S}}^{(m)}, \mathbf{y}) = \prod_{k=1}^{K} p(\boldsymbol{\theta}_k|\mathbf{S}^{(m)}, \boldsymbol{\psi}^{(m)}, \mathbf{y}), \tag{22}$$

with an obvious extension to more than one block.

Consider, for illustration, Gaussian mixtures under the non-conjugate hierarchical prior $p(\boldsymbol{\theta}_k|\boldsymbol{\psi}) = p(\boldsymbol{\mu}_k)p(\boldsymbol{\Sigma}_k|\boldsymbol{\psi})$ where $\boldsymbol{\psi} \sim \mathcal{W}(g_0, \mathbf{G}_0)$ follows a Wishart distribution. In this case, the components in (22) read:

$$q_K(\boldsymbol{\theta}_k|\tilde{\mathbf{S}}^{(m)}, \mathbf{y}) = p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{(m)}, \mathbf{S}^{(m)}, \mathbf{y})p(\boldsymbol{\Sigma}_k|\boldsymbol{\mu}_k^{(m-1)}, \boldsymbol{\psi}^{(m-1)}, \mathbf{S}^{(m)}, \mathbf{y}),$$

hence $\tilde{\mathbf{S}}^{(m)} = (\mathbf{S}^{(m)}, \boldsymbol{\Sigma}_1^{(m)}, \ldots, \boldsymbol{\Sigma}_K^{(m)}, \boldsymbol{\mu}_1^{(m-1)}, \ldots, \boldsymbol{\mu}_K^{(m-1)}, \boldsymbol{\psi}^{(m-1)})$. Note that the marginal prior $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|K)$ can be evaluated as in (21):

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|K) = \frac{p(\boldsymbol{\psi}^\star)}{p(\boldsymbol{\psi}^\star|\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)} \prod_{k=1}^{K} p(\boldsymbol{\mu}_k)p(\boldsymbol{\Sigma}_k|\boldsymbol{\psi}^\star).$$

5 Applications

By combining bridge sampling (BS), importance sampling (IS) and reciprocal importance sampling (RI) with the various ways to construct the importance density, following marginal likelihood estimators are obtained: $\hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$, $\hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$, and $\hat{p}_{\mathrm{RI},F}(\mathbf{y}|K)$ using full permutation sampling (Algorithm 3), where the fully balanced importance density $q_K^F(\boldsymbol{\vartheta})$ is constructed from (10) with M_0 components per mode, as well as $\hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$, $\hat{p}_{\mathrm{IS},D}(\mathbf{y}|K)$, and $\hat{p}_{\mathrm{RI},D}(\mathbf{y}|K)$ using double random permutation sampling (Algorithm 2), where the (nearly) balanced importance density $q_K^D(\boldsymbol{\vartheta})$ is constructed from (9) with $Q = M_0K!$, ensuring that each mode is visited on average M_0 times.

The aim of this section is to apply these marginal likelihood estimators to a wide range of mixture models for increasing values of K and to compare them to the simple random permutation estimators $\hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$, $\hat{p}_{\mathrm{IS},R}(\mathbf{y}|K)$, and $\hat{p}_{\mathrm{RI},R}(\mathbf{y}|K)$ (Frühwirth-Schnatter, 2004) based on the importance density $q_K^R(\boldsymbol{\vartheta})$ defined in Algorithm 1 with $Q = M_0K!$.

Unless stated otherwise, MCMC estimation is performed for a given K for $M = 12{,}000$ draws after a burn-in of 5000. Construction of all importance densities is based on $M_0 = 100$ and the various bridge sampling estimators are based on $L = M = 12{,}000$. All computations are carried out in MATLAB, using the bayesf package (Frühwirth-Schnatter, 2019). Results are visualised by plotting the nine estimators $\log \hat{p}(\mathbf{y}|K)$ as well as $\log \hat{p}(\mathbf{y}|K) \pm 3\mathrm{SE}$ in the order $\log \hat{p}_{\mathrm{BS},\bullet}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{IS},\bullet}(\mathbf{y}|K)$, and $\log \hat{p}_{\mathrm{RI},\bullet}(\mathbf{y}|K)$ over K, where the standard errors SE are computed as in Frühwirth-Schnatter (2004).

5.1 Finite mixture models

Subsequently, finite mixture analysis is based on the prior $\boldsymbol{\eta} \sim \mathcal{D}_K(e_0)$ with $e_0 = 4$ for the weight distribution $\boldsymbol{\eta}$.

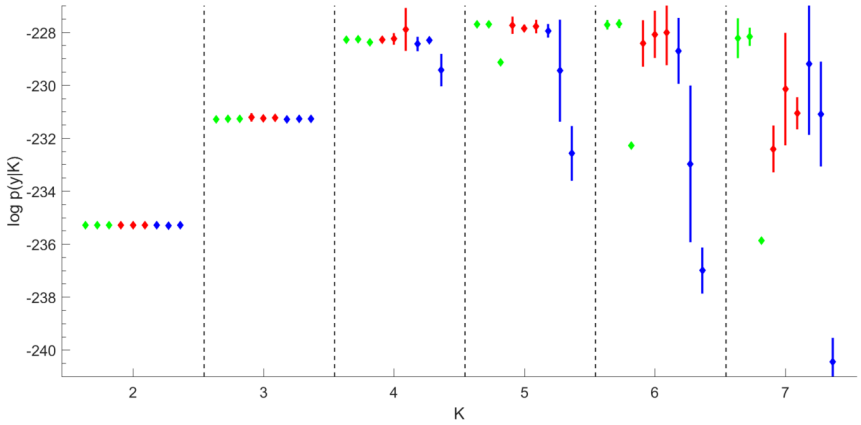

Figure 1 *Marginal likelihood estimation for the* GALAXY DATA *over* $K = 2$ *to* $K = 7$. *For each* K, *nine estimators* $\log \hat{p}_\bullet(\mathbf{y}|K)$ *are given together with* $\log \hat{p}_\bullet(\mathbf{y}|K) \pm 3SE$ *in following order from left to right*: $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$ (*green*); $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ (*dual importance sampling*), $\log \hat{p}_{\mathrm{IS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{IS},R}(\mathbf{y}|K)$ (*red*); $\log \hat{p}_{\mathrm{RI},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},R}(\mathbf{y}|K)$ (*blue*).

5.1.1 *Univariate Gaussian mixtures*. For illustration, marginal likelihoods are computed for univariate Gaussian mixtures $y_i|S_i = k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ for the GALAXY DATA (Richardson and Green, 1997) for $K = 2, \ldots, 7$, using the priors $\mu_k \sim \mathcal{N}(m, R^2)$, $\sigma_k^2 \sim \mathcal{G}^{-1}(2, C_0)$, and $C_0 \sim \mathcal{G}(0.2, 10/R^2)$, where m and R are the midpoint and the length of the observation interval. For a given K, full conditional Gibbs sampling is performed by iteratively sampling from $p(\sigma_k^2|\mu_k, C_0, \mathbf{S}, \mathbf{y})$, $p(\mu_k|\sigma_k^2, \mathbf{S}, \mathbf{y})$, $p(C_0|\sigma_1^2, \ldots, \sigma_K^2)$, $p(\boldsymbol{\eta}|\mathbf{S})$, and $p(\mathbf{S}|\boldsymbol{\vartheta}, \mathbf{y})$, see Frühwirth-Schnatter (2006).

Results of marginal likelihood estimation are visualised in Figure 1. There is a striking difference in the reliability of the nine estimators, in particular as K increases. (Optimal) bridge sampling in combination with the fully symmetric importance density $q_K^F(\boldsymbol{\vartheta})$ and the importance density $q_K^D(\boldsymbol{\vartheta})$ (first two estimators in green) yield the most reliable results. Up to $K = 5$, the dual IS estimator $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ is as good as $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$ and $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$. However, for $K \geq 6$, the standard errors of both bridge sampling estimators are considerably smaller than the standard errors of the dual IS estimator due to their robustness with respect to the tail behaviour of the importance density. Reciprocal importance sampling estimators $\log \hat{p}_{\mathrm{RI},\bullet}(\mathbf{y}|K)$ (in blue) become particularly unreliable as K increases, with extreme bias and huge SE, even for the fully symmetric importance density $q_K^F(\boldsymbol{\vartheta})$.

5.1.2 *Multivariate Gaussian mixtures*. For further illustration, marginal likelihoods are computed for multivariate Gaussian mixtures as in Section 4.1 for the well-known FISHER'S IRIS DATA for $K = 2, \ldots, 5$. We use the normal prior $\boldsymbol{\mu}_k \sim$

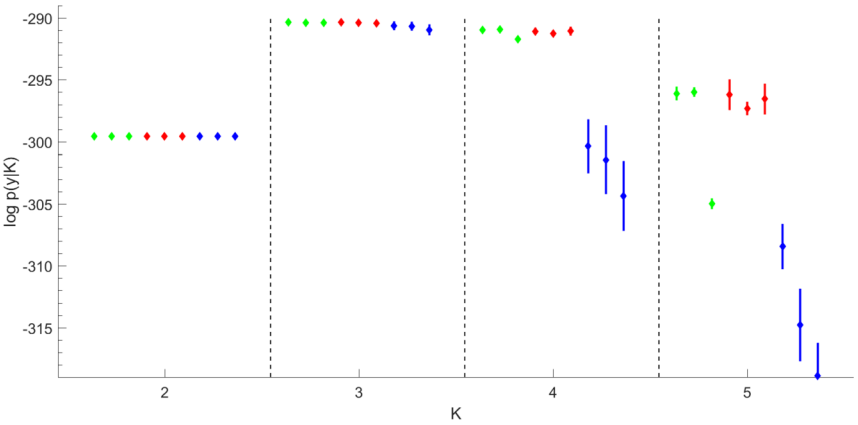

Figure 2 *Marginal likelihood estimation for* FISHER'S IRIS DATA *over* $K = 2$ *to* $K = 5$. *For each* K, *nine estimators* $\log \hat{p}_\bullet(\mathbf{y}|K)$ *are given together with* $\log \hat{p}_\bullet(\mathbf{y}|K) \pm 3SE$ *in following order from left to right*: $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$ (*green*); $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ (*dual importance sampling*), $\log \hat{p}_{\mathrm{IS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{IS},R}(\mathbf{y}|K)$ (*red*); $\log \hat{p}_{\mathrm{RI},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},R}(\mathbf{y}|K)$ (*blue*).

$\mathcal{N}(\mathbf{m}_y, \mathbf{S}_y)$ and the hierarchical inverse Wishart prior $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(c_0, \mathbf{C}_0)$, $\mathbf{C}_0 \sim \mathcal{W}(g_0, g_0/\phi {\mathbf{S}_y}^{-1})$ where $\mathbf{m}_y$ is the componentwise median and $\mathbf{S}_y$ is the sample covariance matrix of the data, $c_0 = 2.5 + (d-1)/2$ and $g_0 = 0.5 + (d-1)/2$, with $d = 4$ being the dimension of the data, and $\phi = (1 - R^2)(c_0 - (d+1)/2)$, where $R^2 = 0.5$ is the amount of explained heterogeneity (Frühwirth-Schnatter, 2006).

Results of marginal likelihood estimation are visualised in Figure 2. Again, (optimal) bridge sampling in combination with the fully symmetric importance density $q_K^F(\boldsymbol{\vartheta})$ and the importance density $q_K^D(\boldsymbol{\vartheta})$ yields the very reliable estimators $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$ and $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$. Up to $K = 4$, the dual IS estimator $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ is as good as these estimators. However, for $K = 5$, the standard errors of the bridge sampling estimators are considerably smaller than the standard errors of the dual IS estimator due to their robustness with respect to the tail behaviour of the importance density. Also for this example, the simple random bridge sampling estimator $\log \hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$ and all reciprocal importance sampling estimators $\log \hat{p}_{\mathrm{RI},\bullet}(\mathbf{y}|K)$ are substantially biased for $K \geq 4$.

5.1.3 *Poisson mixtures*. Finally, marginal likelihoods are computed for Poisson mixtures as in Section 4.1 for the EYE TRACKING DATA (Escobar and West, 1998) for $K = 3, \ldots, 7$ under a Gamma prior with $a_0 = \overline{y}^2/(s_y^2 - \overline{y}^2)$ and $b_0 = a_0/\overline{y}$ (Frühwirth-Schnatter, 2006). Results of marginal likelihood estimation are visualised in Figure 3. Once more, the (optimal) bridge sampling estimators $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$ and $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$ are very precise even for increasing K, whereas all other estimators yield poor results beyond $K = 5$.

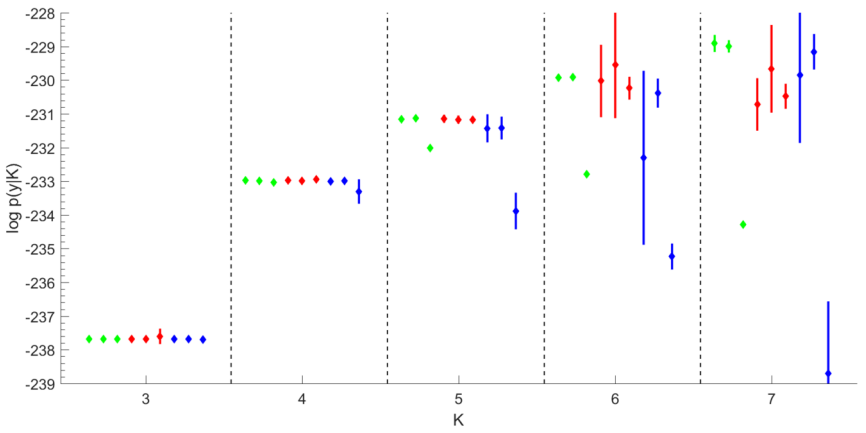

Figure 3 *Marginal likelihood estimation for the* EYE TRACKING DATA *over* $K = 3$ *to* $K = 7$. *For each* K, *nine estimators* $\log \hat{p}_\bullet(\mathbf{y}|K)$ *are given together with* $\log \hat{p}_\bullet(\mathbf{y}|K) \pm 3SE$ *in following order from left to right*: $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$ (*green*); $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ (*dual importance sampling*), $\log \hat{p}_{\mathrm{IS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{IS},R}(\mathbf{y}|K)$ (*red*); $\log \hat{p}_{\mathrm{RI},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},R}(\mathbf{y}|K)$ (*blue*).

5.2 Hidden Markov and Markov switching models for time series analysis

For illustration, we apply the estimators introduced in Section 4.2 to two time series analyzed in Frühwirth-Schnatter (2006). The prior of the transition matrix $\boldsymbol{\xi}$ is defined with $e_p = 4$ and $e_t = 1/(K-1)$ and the initial value S_0 is assumed to follow a uniform distribution.

5.2.1 *Hidden Markov models for the* LAMB DATA. A Markov mixture of Poisson distribution, $y_i|S_i = k \sim \mathcal{P}(\mu_k)$, is applied to the LAMB DATA, a time series of count data (Leroux and Puterman, 1992), under the prior $\mu_k \sim \mathcal{G}(1, 0.5)$. Marginal likelihoods are computed for $K = 2, \ldots, 6$ and visualised in Figure 4.

Using simple random permutation estimators, Frühwirth-Schnatter (2006), p. 353, reports quite unstable estimators of the marginal likelihood for $K = 4$, leading to choose $K = 3$ based on the BS and the RI estimator, whereas the IS estimator indicates $K = 4$. Instability beyond $K = 3$ is also evident in Figure 4, reporting nine different estimators for each K. Also for Markov mixtures, the only reliable estimators of the marginal likelihood are $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$ and $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$ (the first two estimators in green), based on (optimal) bridge sampling in combination with the fully symmetric importance density $q_K^F(\boldsymbol{\vartheta})$ and the importance density $q_K^D(\boldsymbol{\vartheta})$. Up to $K = 4$, the dual IS estimator $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ (first estimator in red) is as good as these estimators.

However, as for finite mixtures, for $K \geq 5$ the standard errors of this estimator are considerably larger than for the two BS estimators due to its lack of robustness with respect to the tail behaviour of the importance density. Once more, reciprocal importance sampling estimators $\log \hat{p}_{\mathrm{RI},\bullet}(\mathbf{y}|K)$ become particularly unreliable

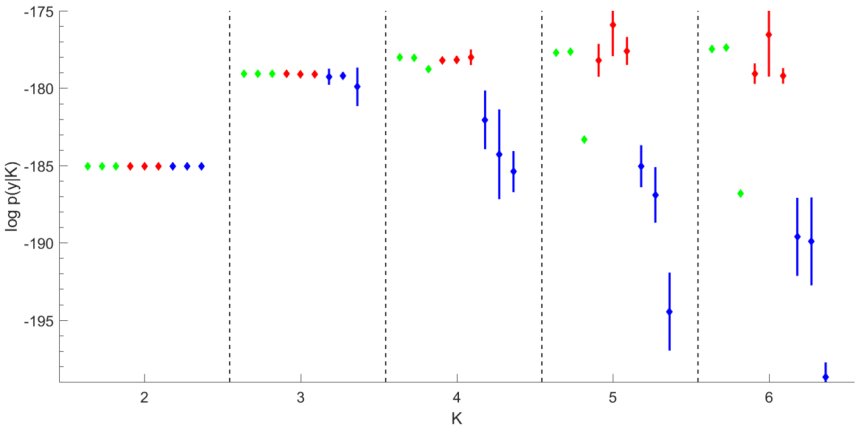

Figure 4 *Marginal likelihood estimation for the* LAMB DATA *data over* $K = 2$ *to* $K = 6$. *For each* K, *nine estimators* $\log \hat{p}_\bullet(\mathbf{y}|K)$ *are given together with* $\log \hat{p}_\bullet(\mathbf{y}|K) \pm 3SE$ *in following order from left to right*: $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$ (*green*); $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ (*dual importance sampling*), $\log \hat{p}_{\mathrm{IS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{IS},R}(\mathbf{y}|K)$ (*red*); $\log \hat{p}_{\mathrm{RI},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},R}(\mathbf{y}|K)$ (*blue*).

as K increases, with extreme bias and huge SE even for the fully symmetric importance density $q_K^F(\boldsymbol{\vartheta})$. The ever increasing marginal likelihood obtained through balanced bridge sampling indicates that the state-specific distribution might be misspecified and more flexible distributions, for example, a negative binomial distribution should be considered.

5.2.2 *Markov switching models for GDP analysis.* A fully Markov switching model of order p with K states is fitted to the GDP DATA as in Frühwirth-Schnatter (2006), assuming that conditional on $S_i = k$,

$$y_i = \delta_{k,1} y_{i-1} + \cdots + \delta_{k,p} y_{i-p} + \zeta_k + \varepsilon_i,$$

where $\varepsilon_i | S_i = k \sim \mathcal{N}(0, \sigma^2_{\varepsilon,k})$. Priors are chosen as $\delta_{k,j} \sim \mathcal{N}(0, 0.25)$, $j = 1, 2$, $\zeta_k \sim \mathcal{N}(0, 10)$, and $\sigma^2_{\varepsilon,k} \sim \mathcal{G}^{-1}(2, 0.5)$.

For illustration, marginal likelihoods are computed for $p = 2$ for $K = 1, \ldots, 4$ and visualised in Figure 5. Also for this Markov switching model, the estimators $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$ and $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$ based on (optimal) bridge sampling in combination with the fully symmetric importance density $q_K^F(\boldsymbol{\vartheta})$ and the importance density $q_K^D(\boldsymbol{\vartheta})$ are very precise, whereas all alternative estimators exhibit considerably larger standard errors and/or considerable bias. The presence of $K = 2$ states is clearly confirmed by this analysis.

5.3 Mixture of experts models in model-based clustering of time series

In many areas of applied statistics, like biometrics, economics, finance, psychometrics, public health, or in social sciences, data are available in the form of panel

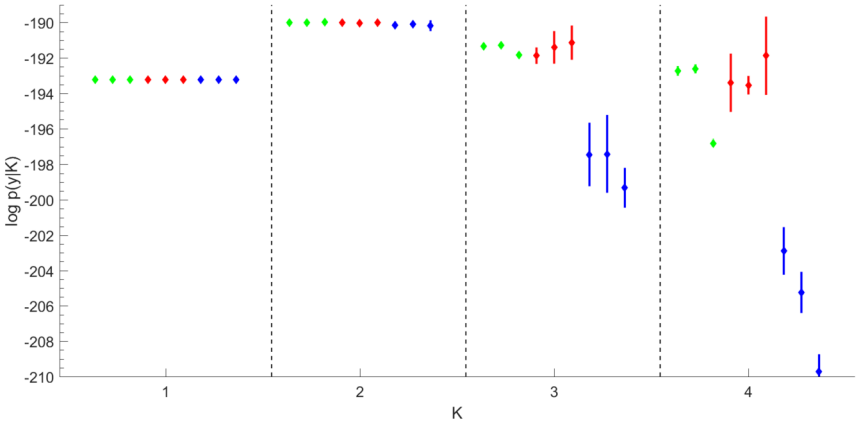

Figure 5 *Marginal likelihood estimation for the* GDP DATA *over* $K = 1$ *to* $K = 4$. *For each* K, *nine estimators* $\log \hat{p}_\bullet(\mathbf{y}|K)$ *are given together with* $\log \hat{p}_\bullet(\mathbf{y}|K) \pm 3SE$ *in following order from left to right*: $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$ (*green*); $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ (*dual importance sampling*), $\log \hat{p}_{\mathrm{IS},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{IS},R}(\mathbf{y}|K)$ (*red*); $\log \hat{p}_{\mathrm{RI},F}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},D}(\mathbf{y}|K)$, $\log \hat{p}_{\mathrm{RI},R}(\mathbf{y}|K)$ (*blue*).

or longitudinal data where, for a given sample of subjects, repeated measurements are taken for a set of variables at several points in time. Standard methods for panel or longitudinal data analysis assume homogeneity across the subjects (Diggle et al., 2002). To capture (unobserved) heterogeneity across subjects, model-based clustering has been applied where each time series is considered to belong to one of K unknown clusters, where each cluster is described by a different data generating mechanism, see, for example, Frühwirth-Schnatter and Kaufmann (2008) and Frühwirth-Schnatter (2011).

To apply model-based clustering, one has to choose the clustering kernel $p(\mathbf{y}_i|\boldsymbol{\theta}_k)$ and the prior class assignment distribution $\Pr(S_i = k|\boldsymbol{\omega})$ for $k = 1, \ldots, K$. To address serial dependence among the observations for each subject, model-based clustering of time series data is often based on dynamic clustering kernels derived from first-order homogeneous or inhomogeneous Markov processes, see Frühwirth-Schnatter (2011) for a review.

Assuming $\Pr(S_i = k|\boldsymbol{\omega}) = \eta_k$ as for finite mixtures would imply that all subjects have the same prior probability to belong to a certain cluster, regardless of their specific characteristics. To achieve more flexibility, covariates $\mathbf{x}_i$ are allowed to influence the weight distribution, modeled as in (16) through a multinomial logit (MNL) model. Such mixture of experts models have been applied to model-based clustering of time series in combination with dynamic regression clustering kernels (Frühwirth-Schnatter and Kaufmann, 2008), Markov chain clustering kernels (Frühwirth-Schnatter et al., 2012), and locally independent MNL clustering kernels (Aßmann and Boysen-Hogrefe, 2011).

A crucial issue in model-based clustering is, of course, how to select the number of clusters present in the panel. Various Bayesian criteria such as marginal likelihoods as well as information based criteria are reviewed in Frühwirth-Schnatter (2011). Below, the various estimators of the marginal likelihood of mixture of experts models introduced in Section 4.4 are applied for model-based clustering of discrete time series arising in panels from the Austrian labour market.

Long-term career outcomes after job loss due to a plant closure—where all workers are automatically displaced—are an often researched topic in labor economics, see, for example, Frühwirth-Schnatter et al. (2018) for the Austrian labour market. Our empirical analysis is based on administrative register data from the Austrian Social Security Database (ASSD), which provides detailed longitudinal information on employment and earnings of all private sector workers in Austria (Zweimüller et al., 2009). To define our sample of displaced workers, we concentrate on all male workers employed during the years 1982 to 1988, who experienced a job displacement due to plant closure in this period. We follow these workers' detailed labor market careers for 4 years prior to job displacement and for 10 years afterwards. We further restrict the sample to workers displaced from firms that have more than 5 employees at least once during the period 1982 to 1988 and who have at least one year of tenure prior to displacement. Moreover, we select workers who were between 35 and 55 years of age at the time of job displacement.

To compare labor market careers after job loss with a counterfactual situation without job displacement, a control group of workers is selected who were employed during the years 1982 to 1988 in firms which did not close down. Following Schwerdt et al. (2010), controls are selected who are very similar to the displaced group in terms of their pre-displacement labor market careers and observable individual characteristics such as age, broad occupation (white versus blue collars) and industry using exact statistical matching. This yields a panel of $N = 17{,}511$ time series, containing 3417 displaced workers and 14,094 controls.

The outcome variable y_{it}, $i = 1, \ldots, N$, $t = 1, \ldots, 40$ is a categorical variable with $J = 7$ categories, among them *employed*, *retired*, *sick*, and *unemployed*. Frühwirth-Schnatter (2011) provides a review how to choose clustering kernels specifically for discrete-valued time series observations, where y_{it} is a categorical variable with J states labelled by $j \in \{1, \ldots, J\}$. Such clustering kernels are based on modelling the probability distribution $\Pr(y_{it} = j|\boldsymbol{\theta}_k)$, $j = 1, \ldots, J$ in terms of class-specific parameters $\boldsymbol{\theta}_k$. If covariate information $\mathbf{w}_{it}$ is available, inhomogeneous Markov chains are used as clustering kernels, by modeling the rows of the transition matrix through a dynamic multinomial logit (MNL) model,

$$\Pr(y_{it} = j|y_{i,t-1} = l, S_i = k) = \frac{\exp(\lambda_{itk,lj})}{\sum_{\tilde{j}=1}^{J} \exp(\lambda_{itk,l\tilde{j}})}, \tag{23}$$

where $\lambda_{itk,lj} = \alpha_{k,lj} + \mathbf{w}_{it}\boldsymbol{\beta}_{k,l}$ depend on the past state $y_{i,t-1} = l$ and cluster specific regression parameter $\boldsymbol{\beta}_{k,j}$ capturing the effect of the covariates $\mathbf{w}_{it}$.

A special version of (23) results, if $\mathbf{w}_{it}$ take on only a few values. Assume, for instance, that the only covariate information to be used for each subject i is a dummy variable g_i. If all H possible combinations $\mathcal{H}_{it} = (y_{i,t-1}, g_i)$ of the past state $y_{i,t-1}$ and the dummy variable g_i are indexed by $h = 1, \ldots, H$, then the dynamic MNL model (23) reduces to a generalized transition matrix $\boldsymbol{\xi}_k$ with H rows. The hth row $\boldsymbol{\xi}_{k,h\cdot} = (\xi_{k,h1}, \ldots, \xi_{k,hJ})$ defines for each cluster $k = 1, \ldots, K$ the conditional distribution of y_{it}, given that the state of the history $\mathcal{H}_{it}$ equals h:

$$\xi_{k,hj} = \Pr(y_{it} = j | \mathcal{H}_{it} = h, S_i = k), \quad j = 1, \ldots, J. \tag{24}$$

Evidently, the clustering kernel for the time series $\mathbf{y}_i$ in state k reads:

$$p(\mathbf{y}_i | \boldsymbol{\xi}_k) = \prod_{h=1}^{H} \prod_{j=1}^{J} \xi_{k,hj}^{N_{i,hj}}, \tag{25}$$

where, for each time series i, $N_{i,hj} = \#\{t \in \{1, \ldots, 40\} | y_{it} = j, \mathcal{H}_{it} = h\}$ is the number of transitions into state j given a history of type h.

Each row $\boldsymbol{\xi}_{k,h\cdot}$ of the generalized transition matrix $\boldsymbol{\xi}_k$ follows a Dirichlet prior distribution,

$$\boldsymbol{\xi}_{k,h\cdot} \sim \mathcal{D}(e_{0,h1}, \ldots, e_{0,hJ}), \quad e_{0,hj} = \max(N_0 \hat{\xi}_{hj}, 0.5),$$

where $N_0 = 2.5$, $\hat{\xi}_{hj} = N_{hj}/N_h$, $N_{hj} = \sum_{i=1}^{N} N_{i,hj}$ is the total number of transitions into state j given a history of type h, and $N_h = \sum_{j=1}^{J} N_{hj}$ is the total number of observations with history h.

To cluster the time series of labor market states, the generalized transition model (25) is used as clustering kernel, with following history $\mathcal{H}_{it}$. It is assumed that the distribution of y_{it} depends on the previous state $y_{i,t-1}$, the broad occupation (blue versus white collar) and the age group (35–44 versus 45–55) of a worker. This yields a transition matrix with $H = 28$ rows. Clustering these data provides quite a challenge due to the high dimensionality both of the data, with a total of $N = 17,511$ time series each with $T = 40$ observations, and the mixture model with a high-dimensional component-specific parameter of dimension $\dim(\boldsymbol{\xi}_k) = 168$.

Marginal likelihoods are computed for two types of mixture of experts models. First, choosing $\mathbf{x}_i \equiv 1$ in (16) corresponds to an alternative parameterization of a finite mixture model. Second, choosing $\mathbf{x}_i = (1 D_i)$, where D_i is 1, iff person i experienced plant closure, and 0 otherwise, assumes that cluster membership depends on whether a person experienced plant closure.

Results of marginal likelihood estimation based on the completely balanced importance density $q_K^F(\boldsymbol{\vartheta})$ are visualised for both models for $K = 2, \ldots, 6$ in Figure 6. All estimators are extremely accurate and bridge and importance sampling yield very similar results. The mixture of experts model clearly dominates the finite mixture model and cluster membership depends on whether a person experienced plant closure or not. For the mixture of experts model, $K = 5$ is selected. The

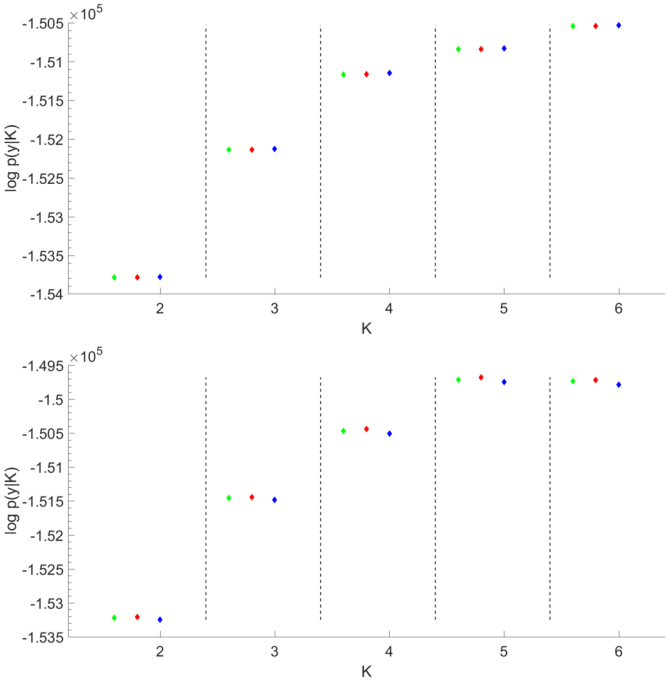

Figure 6 *Marginal likelihood estimation for the plant closure data for a finite mixture model* (*top panel*) *and a mixture of experts model including a plant closure dummy variable* (*bottom panel*) *over* $K = 2$ *to* $K = 6$. *For each* K, *the following estimators including* $\pm 3SE$ *are given*: $\log \hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$ (*green*); $\log \hat{p}_{\mathrm{IS},F}(\mathbf{y}|K)$ (*dual importance sampling*, *red*); $\log \hat{p}_{\mathrm{RI},F}(\mathbf{y}|K)$ (*blue*).

same number of clusters was identified for a closely related data set in Frühwirth-Schnatter et al. (2018), using less formal criteria based on economic interpretability of the resulting clusters as well as AIC and BIC. The resulting clusters are rather similar to the clusters obtained in that paper.

6 Concluding remarks

The present paper shows that sampling-based estimators of the marginal likelihood are prone to be biased under a strongly unbalanced importance density, even for the optimal bridge sampling estimator which is fairly robust to the tail behaviour of the importance density. To address this problem, two bridge sampling estimators are suggested to compute the marginal likelihood for finite mixture models and their extensions to Markov mixture, Markov switching and mixture of experts models. These estimators are based on constructing balanced importance densities from the conditional densities arising during Gibbs sampling.

A particularly stable estimator is obtained, when (optimal) bridge sampling is combined with the perfectly balanced importance density $q_K^F(\boldsymbol{\vartheta})$ yielding the full

permutation bridge sampling estimator $\hat{p}_{\mathrm{BS},F}(\mathbf{y}|K)$. This importance density is derived from considering all possible permutations of the mixture labels for a subset of the conditional densities arising during Gibbs sampling. For the double random permutation bridge sampling estimator $\hat{p}_{\mathrm{BS},D}(\mathbf{y}|K)$, two levels of random permutations are applied, first to permute the labels of the MCMC draws and second to independently permute the labels of the conditional densities arising during Gibbs sampling. The double random permutation estimator provides a simple, yet effective improvement concerning balance and avoids the bias observed for the simple random permutation bridge sampling estimator $\hat{p}_{\mathrm{BS},R}(\mathbf{y}|K)$ for all case studies for larger values of K.

A wide range of applications of these balanced bridge sampling estimators shows very good performance in comparison to importance and, in particular, to reciprocal importance sampling estimators derived from the same importance densities. As the case studies demonstrate, this is true for all types of finite mixtures, including Markov switching and mixture of experts models. The reliability of these estimators results from two main factors. First, these estimators are robust to the tail behaviour of the importance density compared to the mixture posterior. Second, these estimators rely on an importance density that mimics the multimodality of the mixture posterior in a (nearly) balanced way.

For reciprocal importance sampling, considerable bias and large standard errors may occur even if the estimator is based on the perfectly balanced importance density $q_K^F(\boldsymbol{\vartheta})$, as reciprocal importance sampling estimators are particularly sensitive to poor tail behaviour of the importance density and cannot be recommended. As opposed to that under balanced importance densities, importance sampling estimators are in general as reliable as bridge sampling estimators for small values of K. However, with increasing K and for overfitting models, importance sampling estimators tend to be less accurate, even if they are based on the perfectly balanced importance density $q_K^F(\boldsymbol{\vartheta})$, as thinner tails of the importance density compared to the posterior in one mode will be replicated in all other $K! - 1$ modes. As a result, the full permutation bridge sampling estimator is recommended as a default choice for marginal likelihood estimation for finite mixtures as well as Markov switching and mixture of experts models for values of K up to 7.

References

Aßmann, C. and Boysen-Hogrefe, J. (2011). A Bayesian approach to model-based clustering for binary panel probit models. *Computational Statistics & Data Analysis* **55**, 261–279. MR2736553

Berger, J. O. and Jefferys, W. H. (1992). Sharpening Ockham's razor on a Bayesian strop. *American Statistician* **80**, 64–72.

Berkhof, J., van Mechelen, I. and Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica* **13**, 423–442. MR1977735

Celeux, G., Frühwirth-Schnatter, S. and Robert, C. P. (2019). Model selection for mixture models—perspectives and strategies. In *Handbook of Mixture Analysis* (S. Frühwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 117–154. Boca Raton, FL: CRC Press. MR3889692

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321. MR1379473

DiCiccio, T. J., 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

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

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press. MR2049007

Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.), *Lecture Notes in Statistics* **133**, 1–22. Berlin: Springer. MR1630073

Frühwirth-Schnatter, S. (1995). Bayesian model discrimination and Bayes factors for linear Gaussian state space models. *Journal of the Royal Statistical Society, Series B* **57**, 237–246. MR1325388

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* **96**, 194–209. MR1952732

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal* **7**, 143–167. MR2076630

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer. MR2265601

Frühwirth-Schnatter, S. (2011). Panel data analysis—A survey on model-based clustering of time series. *Advances in Data Analysis and Classification* **5**, 251–280. MR2860101

Frühwirth-Schnatter, S. (2019). Applied Bayesian Mixture Modelling. Implementations in MATLAB using the package bayesf (Version 4.0). Available at https://www.wu.ac.at/statmath/faculty-staff/faculty/sfruehwirthschnatter/.

Frühwirth-Schnatter, S., Celeux, G. and Robert, C. P., eds. (2019). *Handbook of Mixture Analysis*. Boca Raton, FL: CRC Press. MR3889980

Frühwirth-Schnatter, S. and Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In *Statistical Modelling and Regression Structures—Festschrift in Honour of Ludwig Fahrmeir* (T. Kneib and G. Tutz, eds.) 111–132. Heidelberg: Physica-Verlag. MR2664631

Frühwirth-Schnatter, S., Frühwirth, R., Held, L. and Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing* **19**, 479–492. MR2565319

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics* **26**, 78–89. MR2422063

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis & Classification* **13**, 33–64. MR3935190

Frühwirth-Schnatter, S., Pamminger, C., Weber, A. and Winter-Ebmer, R. (2012). Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics* **27**, 1116–1137. MR3041877

Frühwirth-Schnatter, S., Pittner, S., Weber, A. and Winter-Ebmer, R. (2018). Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering. *Annals of Applied Statistics* **12**, 1786–1830. MR3852698

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew normal and skew-t distributions. *Biostatistics* **11**, 317–336.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514. MR1278223

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

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–1339. MR1035115

Gormley, I. C. and Frühwirth-Schnatter, S. (2019). Mixture of experts models. In *Handbook of Mixture Analysis* (S. Frühwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 271–307. Boca Raton, FL: CRC Press. MR3889697

Lee, J. E. and Robert, C. P. (2016). Importance sampling schemes for evidence approximation in mixture models. *Bayesian Analysis* **11**, 573–597. MR3472003

Lee, K., Marin, J.-M., Mengersen, K. and Robert, C. (2009). Bayesian inference on mixtures of distributions. In *Perspectives in Mathematical Sciences I: Probability and Statistics* (N. N. Sastry, M. Delampady and B. Rajeev, eds.) 165–202. Singapore: World Scientific. MR2581744

Leroux, B. G. and Puterman, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48**, 545–558.

Malsiner Walli, G., Frühwirth-Schnatter, S. and Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing* **26**, 303–324. MR3439375

Meng, X.-L. and Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association* **91**, 1254–1267.

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

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association* **108**, 1339–1349. MR3174712

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* **59**, 731–792. MR1483213

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. *Springer Series in Statistics*. New York/Berlin/Heidelberg: Springer. MR1707311

Rousseau, J., Grazian, C. and Lee, J. E. (2019). Bayesian mixture models: Theory and methods. In *Handbook of Mixture Analysis* (S. Frühwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 53–72. Boca Raton, FL: CRC Press. MR3889980

Schwerdt, G., Ichino, A., Ruf, O., Winter-Ebmer, R. and Zweimüller, J. (2010). Does the color of the collar matter? Employment and earnings after plant closure. *Economics Letters* **108**, 137–140.

Zweimüller, J., Winter-Ebmer, R., Lalive, R., Kuhn, A., Wuellrich, J.-P., Ruf, O. and Büchi, S. (2009). The Austrian Social Security Database (ASSD). Working Paper 0903, NRN: The Austrian Center for Labor Economics and the Analysis of the Welfare State, Linz, Austria.

Institute for Statistics and Mathematics
Department of Finance, Accounting and Statistics
Vienna University of Economics and Business (WU)
Welthandelsplatz 1
Vienna, 1020
Austria
E-mail: sfruehwi@wu.ac.at