An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions

Panagiotis Papastamoulis and George Iliopoulos*

Department of Statistics and Insurance Science, University of Piraeus,
80 Karaoli & Dimitriou str., 18534 Piraeus, Greece

Abstract

The Label Switching is a well-known problem occuring in MCMC outputs in Bayesian mixture modelling. In this paper we propose a formal solution to this problem by considering the space of the artificial allocation variables. We show that there exist certain subsets of the allocation space leading to a class of nonsymmetric distributions that have the same support with the symmetric posterior distribution and can reproduce it by simply permutating the labels. Moreover, we select one of these distributions as a solution to the label switching problem using the simple matching distance between the artificial allocation variables. The proposed algorithm can be used in any mixture model and its computational cost depends on the length of the simulated chain but not on the parameter space dimension. Real and simulated data examples are provided in both univariate and multivariate settings. Supplemental material for this article is available online.

Keywords: Mixtures of distributions; Markov chain Monte Carlo; label switching problem; data augmentation; Pivotal Reordering algorithm; genuine multimodality.

1 Introduction

Assume that the observed data $\boldsymbol{x} = (x_1, \ldots, x_n)$ is the realization of a random sample from a finite mixture of distributions,

$$X_i \sim f(x|\boldsymbol{p}, \boldsymbol{\theta}) = \sum_{j=1}^{k} p_j f(x; \theta_j), \quad i = 1, \ldots, n, \tag{1}$$

*Corresponding author; e-mail: geh@unipi.gr

where the weights $\boldsymbol{p} = (p_1, \ldots, p_k)$ are positive and sum to one and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ are the component specific parameters and may be either univariate or multivariate quantities such as vectors or matrices. Throughout this paper, the number of components k is assumed to be known. In many cases, it is convenient to assume that each observation x_i has arised from one of the k components, say component $z_i \in \{1, \ldots k\}$. Then, $z_1, \ldots, z_n$ can be considered as realizations of corresponding independent and identically distributed random variables $Z_1, \ldots, Z_n$ with probability mass function $P(Z_i = j|\boldsymbol{p}) = p_j$, $j = 1, \ldots, k$, for $i = 1, \ldots, n$. This means that, conditional on $Z_i = j$, X_i is distributed according to the jth mixture component, $f(x; \theta_j)$. Notice that $z_1, \ldots, z_n$ are unobserved (otherwise (1) would be no longer a mixture of distributions) and so they need to be treated as missing data.

The EM algorithm is a standard frequentist method for estimation in missing data models and is guaranteed to converge to a local maximum of the likelihood. On the other hand, in a Bayesian setting, the Gibbs sampler can be used in order to simulate a Markov chain with the posterior as limit distribution. In the case of mixtures, both approaches fully exploit the missing data structure described above; at each step, the EM algorithm maximizes the complete likelihood function conditional on the expected values of the latent variables while the Gibbs sampler simulates the values of $Z_1, \ldots, Z_n$ from their full conditional posterior distributions.

An important feature of a mixture model is that the likelihood is invariant under permutations of the components' indices. In a Bayesian setup, if the prior information for the parameters (p_j, θ_j) is the same for all component labels $j = 1, \ldots, k$, then the same holds for the posterior distribution as well. In such cases it turns out that the parameters are not identifiable. When MCMC methods are used for simulation from the posterior distribution, this nonidentifiability leads to the so-called label switching phenomenon, the presence of which has both pros and cons. While it serves as a necessary condition for the convergence of the MCMC algorithms, at the same time renders the parameter estimation procedure non-trivial. In the literature many approaches have been proposed to deal with this phenomenon varying from simple artificial identifiability constraints (Diebolt and Robert, 1994, Richardson and Green, 1997, Frühwirth-Schnatter, 2001) to more sophisticated algorithms based on the Kullback-Leibler divergence (Stephens, 1997a, 2000) or on label invariant loss functions (Celeux et al., 2000) yet none of them is both simple and efficient. In what follows, we propose a simple method that operates on the space of the latent variables

$z_1, \dots, z_n$ and succesfully solves the label switching problem. It has many advantages compared to previous approaches; in particular, it requires small amount of computational effort and is not affected by the dimensionality of the parameter space. Moreover, in the case where the posterior distribution exhibits genuine multimodality, the succesful solution of the label switching problem allows to efficiently post-processing the reordered output by some standard clustering algorithm (e.g. the K-means clustering algorithm) in order to reveal all minor modes provided that they have been explored by the original sampler. Another advantage of the succesful solution is that it leads to better estimates for the parameters that can be used directly in order to obtain a good plug-in density estimate of the posterior distribution.

The rest of the paper is organized as follows. In Section 2, previous approaches for solving the label switching phenomenon are briefly described. The Equivalence Classes Representatives (ECR) algorithm is introduced in Section 3. The approach is justified by providing a rather new expression of the posterior distribution as an equally weighted mixture, and showing that the algorithm produces a sequence that converges to one of this mixture's components. Furthermore, by using ideas from the Pivotal Reordering algorithm of Marin et al. (2005), a practical implementation of ECR algorithm is suggested. In Section 4, the perfomance of the method is compared with that of previous ones in both univariate and multivariate settings. The paper concludes in Section 5 with a discussion. An appendix containing proofs and useful lemmas can be found online as supplemental material.

2 Label switching phenomenon and previous solutions

Let $\mathcal{T}_k$ be the set of permutations of the component indices $\{1, \dots, k\}$. For some $\tau = (t_1, \dots, t_k) \in \mathcal{T}_k$ consider the corresponding permutation of the parameter vector $\tau(\boldsymbol{p}, \boldsymbol{\theta}) = (p_{t_1}, \dots, p_{t_k}, \theta_{t_1}, \dots, \theta_{t_k})$. The root of the label switching phenomenon is the fact that the likelihood $L(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n}\{p_1 f(x_i; \theta_1) + \dots + p_k f(x_i; \theta_k)\}$ is invariant with respect to the permutations of the component labels as it is obvious that $L(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = L(\tau(\boldsymbol{p}, \boldsymbol{\theta})|\boldsymbol{x})$, $\forall \tau \in \mathcal{T}_k$ (see Stephens, 2000).

In a Bayesian analysis, if the prior distribution does not distinguish the component parameters between each other (which is the most common case), then the resulting posterior distribution will be invariant in the permutations of the labels, since it will be proportional to the product of a symmetric likelihood with a symmetric prior distribution. In other words, the parameters are not marginally identifiable as their marginal distributions are exactly the same. Hence, if a sample is simulated from the posterior distribution, the standard method of ergodic averages for estimating the weights and the component specific parameters will lead to nonsensible estimates as they will be the same for every mixture component.

It is desirable for MCMC algorithms to properly explore the posterior distribution, and what we can at least demand is the presence of the label switching phenomenon. However, it is well-known that the Gibbs sampler rarely switches between the symmetric modes. In situations where no label switching occurs, proper label switching moves can be incorporated to guarantee the presence of the phenomenon (see Papaspiliopoulos and Roberts, 2008). On the other hand at the same time we need a simple method to “undo” the label switching in order to derive proper estimates and this can be done by applying suitable permutations to the simulated values. It turns out that this is equivalent to choosing one of the symmetric modes and switching all simulated values to this particular one.

Below, we briefly review some known approaches which attempt to solve the label switching problem.

2.1 Artificial Identifiability Constraints

An identifiability constraint (IC) is a condition on the parameter space of $(\boldsymbol{p}, \boldsymbol{\theta})$ which is satisfied by only one permutation of the parameters. ICs were used, among others, by Diebolt and Robert (1994) and Richardson and Green (1997). They have come under strong criticism in the literature (see for example Celeux, 1997, Celeux et al., 2000, and Stephens, 1997a, 1997b, 2000). One problem with this approach is the choice of the constraint. A more general difficulty of using ICs occurs in multivariate problems. Moreover there are situations where the posterior distribution is genuinely multimodal and no IC can isolate both its main and minor modes succesfully (see e.g. Grün and Leisch, 2009, Section 6.1).

An alternative approach was provided by Frühwirth-Schnatter (2001) who used a random permutation sampler (RPS) in order to ensure that all $k!$ symmetric modes have been visited. Frühwirth-Schnatter then applied exploratory data analysis on a preliminary MCMC run from the RPS in order to find suitable identifiability constraints that separate the components between them. Afterwards, a constrained permutation sampler is used to produce a sample from the constrained posterior distribution, that is, the posterior distribution constrained to the subset of the parameter space that satisfies the artificial IC chosen on the previous step. Nevertheless, the previous drawbacks are present in this case as well.

2.2 Pivotal Reordering Algorithm

A simple method to undo the label switching without imposing an identifiability constraint is the Pivotal Reordering algorithm introduced by Marin et al. (2005) (see also Marin and Robert, 2007). The Monte Carlo approximation of the Maximum A Posteriori (MAP) estimate, i.e., the simulated value that maximizes the posterior distribution, is used as a pivot to reorder all simulated points by simply minimizing a certain distance in the parameter space. In the case of euclidean distance, this task is equivalent to the maximization of the canonical scalar product. The approach works well in simple cases, but in cases of genuine multimodality it has some drawbacks due to the inability of the MAP estimate to accomodate competing explanations of the data (cf. Jasra et al., 2005). This is illustrated via some examples in Section 4.

2.3 Kullback–Leibler divergence based algorithms

Stephens (1997a, 2000) developed an algorithm that makes the permuted sample points to agree as much as possible on the $n \times k$ matrix of classification probabilities $\pi_{ij} = p_j f(x_i|\theta_j)/\sum_{l=1}^{k} p_l f(x_i|\theta_l)$. Stephens measures the distance between two matrices of classification probabilities $\Pi = (\pi_{ij})$ and $Q = (q_{ij})$ using the Kullback–Leibler (KL) divergence $D(\Pi||Q) = \sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij} \log \frac{\pi_{ij}}{q_{ij}}$. Based on a simulated output of length M, the algorithm finds suitable permutations τ_t, $t = 1, \ldots, M$, and a matrix of classification probabilitites $\widehat{\Pi}$ in order to minimize $\mathcal{D} = \sum_{t=1}^{M} D(\tau_t \Pi^{(t)}||\widehat{\Pi})$. As Stephens notes, this algorithm may be computationally quite demanding in memory. Recently, Grün and Leisch (2009) proposed a relabelling and clustering approach extending the algorithm of Stephens to cases where genuine multimodality takes place. More specifically, they introduced a method where the mode allocations and the relabelling of components are silmutaneously determined.

2.4 Label Invariant Loss Functions

A fully decision theoretic approach has been introduced by Celeux et al. (2000) and applied also by Hurn et al. (2003). The method proceeds by defining a loss function that is invariant to the labelling and then minimizing the posterior expected loss. Typically, the minimization step cannot be performed analytically, and so stochastic implementation methods (e.g., simulated annealing) should be implemented.

From a Bayesian point of view, this method is more satisfactory than the previous ones since inference is drawn conditional solely on the data. On the other hand, its main drawback is the high computational cost. A second drawback is the fact that the minimization of the posterior expected risk may not be always feasible restricting possibly the applicability of the method to a class of loss functions that may not make sense for the decision problem at hand (see Jasra, et al., 2005).

The overall message is that there is no solution to the label switching problem that is both simple and efficient to be applied in general settings. Therefore, a method that succesfully solves the problem and requires little computational effort is needed.

3 The Equivalence Classes Representatives Algorithm

In this section we describe a simple yet efficient method for solving the label switching problem. In order to justify it, a deeper insight to the posterior distribution will be useful.

3.1 A mixture representation of the posterior distribution

Note that the posterior distribution of $(\boldsymbol{p}, \boldsymbol{\theta})$ can be expressed as

$$f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = \sum_{\boldsymbol{z}\in\mathcal{Z}} w(\boldsymbol{z}|\boldsymbol{x}) f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) \tag{2}$$

where $w(\boldsymbol{z}|\boldsymbol{x})$ denotes the posterior weight of the allocation vector $\boldsymbol{z}$, and $f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z})$ denotes the posterior distribution of $(\boldsymbol{p}, \boldsymbol{\theta})$ given $(\boldsymbol{x}, \boldsymbol{z})$ (see Marin et al., 2005). In what follows, everything is based on permutations of the allocations. So, in order to be strict we give the following definition.

Definition 3.1 *Let $\tau = (t_1, \ldots, t_k) \in \mathcal{T}_k$, be a permutation of the index set. The corresponding relabelling $\tau\boldsymbol{z}$ of the allocation vector $\boldsymbol{z} = (z_1, \ldots, z_n) \in \mathcal{Z} := \{1, \ldots, k\}^n$ is given by $\tau\boldsymbol{z} = (t_{z_1}, \ldots, t_{z_n}) \in \mathcal{Z}$.*

Note that when the components are not labelled, it holds $w(\tau\boldsymbol{z}|\boldsymbol{x}) = w(\boldsymbol{z}|\boldsymbol{x})$ for all $\tau \in \mathcal{T}_k$ and $\boldsymbol{z} \in \mathcal{Z}$, and this results in a posterior distribution that is symmetric with respect to the permutations of the labels. It is easy to see that Definition 3.1 implies an equivalence relation on the allocation space $\mathcal{Z}$.

Definition 3.2 *Two allocation vectors $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}$ will be said to be equivalent if there exists $\tau \in \mathcal{T}_k$ such that $\boldsymbol{z}_1 = \tau\boldsymbol{z}_2$.*

Let $\Xi_{\boldsymbol{z}} = \{\tau\boldsymbol{z} : \tau \in \mathcal{T}_k\}$ denote the equivalence class of $\boldsymbol{z} \in \mathcal{Z}$. It is easy to see that $\Xi_{\boldsymbol{z}}$ contains $k!/(k - k_0(\boldsymbol{z}))!$ elements, where $k_0(\boldsymbol{z})$ denotes the number of nonempty components for a given allocation vector $\boldsymbol{z}$. Note that using an inclusion–exclusion argument, it can be concluded that the total number of classes (for given k and n) equals $\sum_{i=1}^{k}(i^n/i!)\sum_{j=0}^{k-i}(-1)^j/j!$. Consider now an arbitrary set $\mathcal{Z}_0$, consisting of exactly one representative from each equivalence class and let

$$f_{\mathcal{Z}_0}(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) := \sum_{\boldsymbol{z} \in \mathcal{Z}_0} \frac{k!\, w(\boldsymbol{z}|\boldsymbol{x})}{(k - k_0(\boldsymbol{z}))!} f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}). \tag{3}$$

It is easy to verify that the weights $k!\, w(\boldsymbol{z}|\boldsymbol{x})/(k - k_0(\boldsymbol{z}))!$ sum to one when $\boldsymbol{z} \in \mathcal{Z}_0$. Therefore, $f_{\mathcal{Z}_0}(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ is a probability density function with the same support as $f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$, since it is a mixture of the distributions $f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z})$, for $\boldsymbol{z} \in \mathcal{Z}_0$. Notice here that it differs from the conditional posterior distribution of $(\boldsymbol{p}, \boldsymbol{\theta})$ given $\boldsymbol{z} \in \mathcal{Z}_0$. Observe also that $f_{\mathcal{Z}_0}$ is nonsymmetric for any choice of $\mathcal{Z}_0$. Moreover, it holds the following.

Lemma 3.1 *The posterior distribution of $(\boldsymbol{p}, \boldsymbol{\theta})$ can be expressed as*

$$f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = \frac{1}{k!} \sum_{\tau \in \mathcal{T}_k} f_{\mathcal{Z}_0}\left(\tau(\boldsymbol{p}, \boldsymbol{\theta})|\boldsymbol{x}\right) \tag{4}$$

for any choice of the equivalence classes representatives $\mathcal{Z}_0$.

Lemma 3.1 shows that the posterior distribution of $(\boldsymbol{p}, \boldsymbol{\theta})$ can be written as an equally weighted mixture of nonsymmetric distributions. This representation has a fruitful interpretation in terms of the label switching phenomenon. First, we conclude that a sample from the distribution $f_{\mathcal{Z}_0}(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ is sufficient to produce a sample from the posterior distribution $f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ and this can be done simply by permuting the labels of the simulated values, as expression (4) shows. More important is the reverse direction. Assume for convenience that the sampler uses data augmentation. (This is actually not a restriction since in the opposite case the output can be afterwards augmented as described in Section 5.) If we have obtained a sample from the symmetric posterior distribution, we are able to produce a sample from $f_{\mathcal{Z}_0}(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ according to the following scheme: Suppose that at the ith iteration of the original sampler we simulate $(\boldsymbol{z}^{(i)}, (\boldsymbol{p}, \boldsymbol{\theta})^{(i)})$ with $\boldsymbol{z}^{(i)} \in \mathcal{Z}$. Let $\tau_i \in \mathcal{T}_k$ be such that $\tau_i \boldsymbol{z}^{(i)} \in \mathcal{Z}_0$. Obviously, it holds $\boldsymbol{z}^{(i)} = \tau_i^{-1}\tau_i \boldsymbol{z}^{(i)}$. Then, by Lemma A.1 in the Appendix we get

$$f(\boldsymbol{z}^{(i)}, (\boldsymbol{p}, \boldsymbol{\theta})^{(i)}|\boldsymbol{x}) = f(\tau_i \boldsymbol{z}^{(i)}, \tau_i^{-1}(\boldsymbol{p}, \boldsymbol{\theta})|\boldsymbol{x}). \tag{5}$$

The last expression is the key in order to obtain a sample from $f_{\mathcal{Z}_0}(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ while we have an (augmented) sample from the posterior distribution of $(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta})$, as shown in Lemma A.2 in the Appendix. Moreover, by setting $\mathcal{T}_{\boldsymbol{z}} := \{\tau \in \mathcal{T}_k : \tau\boldsymbol{z} \in \mathcal{Z}_0\}$ for any $\boldsymbol{z} \in \mathcal{Z}$, we have the following convergence result.

Proposition 3.1 *Let $(\boldsymbol{z}^{(i)}, (\boldsymbol{p}, \boldsymbol{\theta})^{(i)})$, $i = 1, 2, \ldots$ be a Markov chain with limit distribution $f(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$. For all i let τ_i be uniformly distributed on $\mathcal{T}_{\boldsymbol{z}^{(i)}}$. Then the limit distribution of the sequence $\tau_i^{-1}(\boldsymbol{p}, \boldsymbol{\theta})^{(i)}$ is $f_{\mathcal{Z}_0}(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$.*

By the construction of the set $\mathcal{Z}_0$, $f_{\mathcal{Z}_0}$ is not symmetric and can produce a sample from the uncostrained posterior distribution taking into account all permutations of the component indices. Hence, $f_{\mathcal{Z}_0}$ can be used in order to solve the label switching phenomenon. Moreover, this distribution has the same support as the (original) posterior distribution. (This is not the case for the constrained posterior distributions obtained when we adopt the permutation sampler of Frühwirth-Schnatter, 2001, or the Pivotal Reordering algorithm of Marin et al., 2005.) With that in mind, our method for solving the label switching problem proceeds as follows.

The Equivalence Classes Representatives (ECR) Algorithm

```
1. Determine $\mathcal{Z}_0$ and obtain a simulated output $(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta})^{(i)}$, $i = 1, \ldots, M$, with target distribution $f(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta} | \boldsymbol{x})$.
2. For $i = 1, \ldots, M$:
    (a) Select randomly some permutation $\tau_i \in \mathcal{T}_{\boldsymbol{z}^{(i)}}$.
    (b) Set $(\boldsymbol{z}', \boldsymbol{p}', \boldsymbol{\theta}')^{(i)} = \big(\tau_i \boldsymbol{z}^{(i)}, \tau_i^{-1}(\boldsymbol{p}, \boldsymbol{\theta}^{(i)})\big)$.
3. Use the reordered values $(\boldsymbol{p}', \boldsymbol{\theta}')^{(i)}$, $i = 1, \ldots, M$, in order to approximate the posterior distributions of the weights and of the component specific parameters.
```

3.2 Choosing a set of representatives

In order to determine the set of representatives $\mathcal{Z}_0$, we must clarify aspects like the difference between two sets of representatives, what is a “good” set of representatives, and how such a set can be constructed.

As is apparent from (3), the distribution $f_{\mathcal{Z}_0}$ of the reordered sample of $(\boldsymbol{p}, \boldsymbol{\theta})$ is a mixture of the conditional distributions $f(\boldsymbol{p}, \boldsymbol{\theta} | \boldsymbol{x}, \boldsymbol{z})$, $\boldsymbol{z} \in \mathcal{Z}_0$. Recall that the weights of this mixture are not affected by the selection of the representative, since $w(\tau \boldsymbol{z} | \boldsymbol{x})/(k - k_0(\tau \boldsymbol{z}))!$ is constant with respect to $\tau \in \mathcal{T}_k$. Hence, two different sets of representatives $\mathcal{Z}_0$ and $\mathcal{Z}_0'$ yield two distributions $f_{\mathcal{Z}_0}$ and $f_{\mathcal{Z}_0'}$ that differ only with respect to their components $f(\boldsymbol{p}, \boldsymbol{\theta} | \boldsymbol{x}, \boldsymbol{z})$, $\boldsymbol{z} \in \mathcal{Z}_0$, and $f(\boldsymbol{p}, \boldsymbol{\theta} | \boldsymbol{x}, \boldsymbol{z})$, $\boldsymbol{z} \in \mathcal{Z}_0'$. Observe also that if $\mathcal{Z}_0 \neq \tau \mathcal{Z}_0'$ for all $\tau \in \mathcal{T}_k$, then $f_{\mathcal{Z}_0}$ and $f_{\mathcal{Z}_0'}$ typically have completely different shapes and the first can not be reproduced from the second by applying simply a permutation to $(\boldsymbol{p}, \boldsymbol{\theta})$.

An efficient solution to the label switching problem must not just break the symmetry of the posterior distribution but further eliminate as much as possible the influence of all but one of its symmetric copies. Now, while for any $\mathcal{Z}_0$ the distribution $f_{\mathcal{Z}_0}$ is nonsymmetric, an arbitrary selection of the equivalence classes representatives would retain a significant portion of some symmetric copies’ magnitude. In order to avoid that, the representatives must be selected in such a way so that all components of $f_{\mathcal{Z}_0}$ have their masses concentrated as much as possible to one of the $k!$ symmetric high posterior density areas. Clearly, the areas at which the components concentrate their masses depend on the corresponding allocation vectors $\boldsymbol{z}$. Moreover, “similar”

$\boldsymbol{z}$'s will give rise to components which are close to each other. A natural measure of the similarity of two allocation vectors is the number of their matching allocations:

Definition 3.3 *The S similarity measure of two allocation vectors $\boldsymbol{z}_1 = (z_{11}, \ldots, z_{1n})$ and $\boldsymbol{z}_2 = (z_{21}, \ldots, z_{2n})$ is defined by $S(\boldsymbol{z}_1, \boldsymbol{z}_2) := \sum_{i=1}^{n} I(z_{1i} = z_{2i})$, where $I(A)$ is the indicator function of A.*

Obviously, it holds $0 \leq S(\boldsymbol{z}_1, \boldsymbol{z}_2) \leq n$. Moreover, notice that $n - S(\boldsymbol{z}_1, \boldsymbol{z}_2)$ is the simple matching distance between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. Note that although more sophisticated measures could be used, S similarity is perfect for our purposes.

Intuition probably suggests that, in constructing $\mathcal{Z}_0$, all representatives (and, by analogy, the components of $f_{\mathcal{Z}_0}$) should be chosen to be as much similar as possible to each other. However, we are interested in switching everything to a particular high posterior density area rather than in the overall components' similarity. This can be done by simply selecting an appropriate $\boldsymbol{z}^* \in \mathcal{Z}$, include it to $\mathcal{Z}_0$, and choosing as representatives the allocations that are most similar to it.

Under this point of view, the determination of $\mathcal{Z}_0$ reduces to finding a "good" allocation vector $\boldsymbol{z}^*$ which will act as a pivot for the rest classes' representatives and so, the ECR algorithm can be seen as a modification of the Pivotal Reordering algorithm of Marin et al. (2005) (see Subsection 2.2) on the set $\mathcal{Z}$. More specifically, suppose that we want to select the representative $\tau\boldsymbol{z}$ of some equivalence class $\Xi_{\boldsymbol{z}}$. Define $\tau(\Xi_{\boldsymbol{z}}) = \arg\max_{\tau \in \mathcal{T}_k} S(\tau\boldsymbol{z}, \boldsymbol{z}^*)$ to be the permutation which makes $\boldsymbol{z}$ as much similar as possible to $\boldsymbol{z}^*$ and select $\tau(\Xi_{\boldsymbol{z}})\boldsymbol{z}$ as the corresponding representative to be included in $\mathcal{Z}_0$. In the case where there are more than one maximizing permutations, $\tau_1, \tau_2, \ldots$, order $\tau_1\boldsymbol{z}, \tau_2\boldsymbol{z}, \ldots$ lexicographically and choose as $\tau(\Xi_{\boldsymbol{z}})$ the permutation corresponding to the first of them. Hence, the set of representatives is given by $\mathcal{Z}_0 = \cup\{\tau(\Xi_{\boldsymbol{z}})\boldsymbol{z};\ \boldsymbol{z} \in \mathcal{Z}\}$. Notice the difference between the set of possible $\tau(\Xi_{\boldsymbol{z}})$'s and $\mathcal{T}_{\boldsymbol{z}}$ defined in the previous subsection. The former serves for defining the particular $\mathcal{Z}_0$ whilst the latter has a meaning only *after* $\mathcal{Z}_0$ has been defined. On the other hand, the two sets are connected with $\mathcal{T}_{\boldsymbol{z}}$ being a subset of the set of possible $\tau(\Xi_{\boldsymbol{z}})$'s.

3.3 Reordering a simulated output

It is evident that the posterior distribution $f(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ has multiple modes as well. Indeed, (5) implies that if $(\boldsymbol{z}^{(\mathrm{MAP})}, (\boldsymbol{p}, \boldsymbol{\theta})^{(\mathrm{MAP})}) \equiv (\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta})^{(\mathrm{MAP})}$ is a mode, then $(\tau\boldsymbol{z}^{(\mathrm{MAP})}, \tau^{-1}(\boldsymbol{p}, \boldsymbol{\theta})^{(\mathrm{MAP})})$ is also a mode for any permutation τ. Then, an excellent choice for the pivotal value $\boldsymbol{z}^*$ would be $\boldsymbol{z}^{(\mathrm{MAP})}$. This is justified by (2) and the fact that only few allocations have non-negligible posterior weight (cf. Casella et al., 2004). However, the computational effort needed for the analytical evaluation of the modes increases rapidly with the sample size and thus it becomes prohibitive.

Let $(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta})^{(i)}$, $i = 1, \ldots, M$, be a simulated output with target distribution $f(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$. Let also $\hat{\boldsymbol{z}}^{(\mathrm{MAP})}$ be the allocation vector that corresponds to the Monte Carlo approximation of the Maximum a Posteriori estimator $(\hat{\boldsymbol{z}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}})^{(\mathrm{MAP})} = \arg\max_{1 \leq i \leq M} f((\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta})^{(i)}|\boldsymbol{x})$. Since $(\hat{\boldsymbol{z}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}})^{(\mathrm{MAP})}$ consistenly estimates a mode of $f(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$, as the number of iterations M increases $S(\hat{\boldsymbol{z}}^{(\mathrm{MAP})}, \tau\boldsymbol{z}^{(\mathrm{MAP})})$ converges to n for some $\tau \in \mathcal{T}_k$. However, by the discreteness of the allocation vector, the S similarity of $\hat{\boldsymbol{z}}^{(\mathrm{MAP})}$ and $\tau\boldsymbol{z}^{(\mathrm{MAP})}$ (for some $\tau \in \mathcal{T}_k$) will become sooner or later equal to n. But recall that all we need in order to apply our approach is a good pivot $\boldsymbol{z}^*$ and not necessarily the "best" one. So, $\hat{\boldsymbol{z}}^{(\mathrm{MAP})}$ is a satifactory choice as well.

With $\hat{\boldsymbol{z}}^{(\mathrm{MAP})}$ as the pivot, the scenario of existing two (or more) different members of a class maximizing the S similarity measure (and hence having to choose according to the lexicographical order) is quite rare. In particular, this has never occured in the various examples we tried. Note also that for any $\boldsymbol{z} \in \mathcal{Z}$ it holds $\mathcal{T}_{\boldsymbol{z}} = \{\tau = \arg\max_{\tau \in \mathcal{T}_k} S(\tau\boldsymbol{z}, \hat{\boldsymbol{z}}^{(\mathrm{MAP})})\}$.

Besides the MAP estimate, other valid choices for the pivot are the most probable allocation and the allocation vector corresponding to the maximum of the complete likelihood. Moreover, since $\mathcal{Z}_0$ can be *any* set of representatives, one can use as a pivot any allocation vector that has been frequently visited by the MCMC algorithm, provided that n is not very small. In fact, when we tried these different pivot choices we obtained almost identical results. Finally, we underline that we avoid more complicated schemes for the determination of $\mathcal{Z}_0$ because the selection of the pivot $\boldsymbol{z}^*$ has not the drawbacks of the Pivotal Reordering algorithm. Since we are dealing with the space of artificial allocation variables rather than the parameter space, we can take advantage of its discrete nature and the small number of the allocations with non-negligible posterior weight. The latter implies that the majority of classes have almost zero weight and thus they do not contribute much to the posterior distribution.

4 Examples

In this section we illustrate our approach via both univariate and multivariate datasets. The first example shows analytically how the method transforms the posterior distribution. Afterwards, the method is illustrated via simulated and real datasets while at the same time the results are compared with those obtained by other approaches as the Pivotal Reordering and KL based algorithms. In all cases the number of components is assumed to be known. Note that in all MCMC algorithms a label switching move is added (see Papaspiliopoulos and Roberts, 2008) in order to ensure the presence of the label switching phenomenon. Finally, the reported standard errors have been estimated by running the same sampler 100 times independently with different starting values. All simulations and reorderings have been performed on a Pentium IV using Fortran 90. The optimal permutations were found using Carpaneto's (1980) Fortran routine for solving the assignment problem.

4.1 An exact illustration of the proposed method

We simulated $\boldsymbol{x} = (6, 12, 9, 4, 6)$ from a mixture of two Poisson distributions with known and equal weights, $0.5\mathcal{P}(\theta_1) + 0.5\mathcal{P}(\theta_2)$, where $\boldsymbol{\theta} = (\theta_1, \theta_2) = (5, 7)$. We assumed further that θ_1, θ_2 are a priori independent with the same prior distribution $\mathcal{G}(1.2, 0.2)$, that is, gamma with mean $1.2/0.2 = 6$. The resulting symmetric posterior distribution of $\theta_1, \theta_2 | \boldsymbol{x}$ is shown in Figure 1(a). Using `Mathematica` we found its two symmetric modes at $(7.75, 6.01)$ and $(6.01, 7.75)$ (indicated by arrows). In Figures 1(b) and 1(c) we have also plotted the (nonsymmetric) distributions $f_{\mathcal{Z}_0}(\boldsymbol{\theta}|\boldsymbol{x})$ for two different choices of $\mathcal{Z}_0$. In the first case, the classes' representatives have been randomly selected whereas in the second case each representative is chosen to be as much similar as possible to $\boldsymbol{z}^* = (1, 2, 2, 1, 1)$ which is the allocation vector corresponding to the maximum of the $f(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{x})$. Clearly, while in both cases the resulting distributions break the symmetry of the posterior, the latter should be preferred since the magnitude of the symmetric mode has been totally vanished in contrast to the former where a significant portion of the symmetric mode is retained.

In general, the apparent modes of a mixture of distributions may be far apart from the modes of its components. Indeed, in Figure 1(c) we can see that the mode of $f_{\mathcal{Z}_0}$ is at $(5.35, 8.36)$. Moreover, $f_{\mathcal{Z}_0}$ exhibits a minor mode at $(7.15, 1.00)$ which is not visible in the posterior distribution. Its appearence is due to the fact that the posterior probability of an empty component is sufficiently

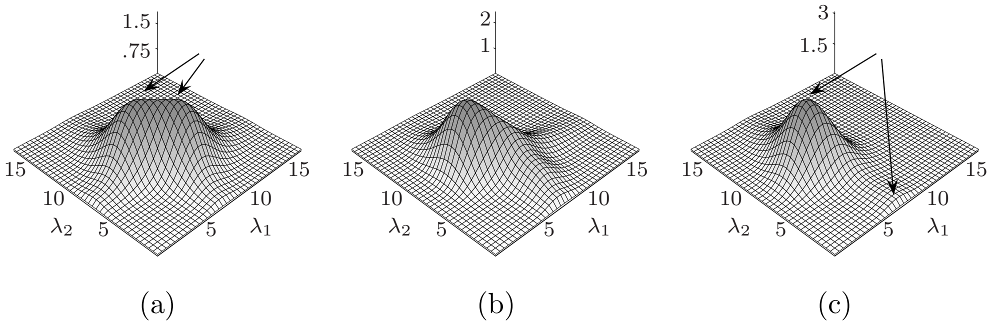

Figure 1: (a) The symmetric posterior distribution of $\theta_1, \theta_2|\boldsymbol{x}$. (b,c) The distributions $f_{\mathcal{Z}_0}(\boldsymbol{\theta}|\boldsymbol{x})$ when the equivalence classes representatives are selected randomly and by maximizing the S similarity to $\boldsymbol{z}^* = (1, 2, 2, 1, 1)$, respectively. (All densities are shown up to the same multiplicative constant.)

large: straightforward calculation yields $w(1, 1, 1, 1, 1|\boldsymbol{x}) = w(2, 2, 2, 2, 2|\boldsymbol{x}) \approx .0394$. This is further justified by the value of the second coordinate of the minor mode which is actually the mode of the prior. Note that when we explored the posterior distribution using the Gibbs sampler the results produced by the ECR algorithm totally agreed with the theoretical ones. In particular, the reordered output explores the minor mode at the correct rate; the weight assigned by a K-means clustering algorithm to the corresponding cluster was approximately 7.97%, that is, almost twice the weight of an empty component as expected.

4.2 ECR Algorithm versus Pivotal Reordering Algorithm

In order to illustrate the differences between the ECR algorithm and the standard Pivotal Reordering algorithm of Marin et al. (2005) we simulated data from two mixtures of normal distributions, namely,

$$0.10\mathcal{N}(-20, 1) + 0.65\mathcal{N}(20, 3) + 0.25\mathcal{N}(21, 0.5), \tag{6}$$

$$0.20\mathcal{N}(19, 5) + 0.20\mathcal{N}(19, 1) + 0.25\mathcal{N}(23, 1) + 0.20\mathcal{N}(29, 0.5) + 0.15\mathcal{N}(33, 2). \tag{7}$$

From (6) we simulated $n = 160$ observations while from (7) we simulated $n = 600$ observations. In both cases we used the random beta model of Richardson and Green (1997) but with the number of components fixed at their true values. Afterwards, the simulated samples (after burn-in) were reordered according to both the Pivotal Reordering and ECR algorithms. In the first two rows of Figure 2 we plot the reordered raw values of the means for the two methods as well as the data histograms together with the corresponding plug-in density estimates. Moreover, the resulting

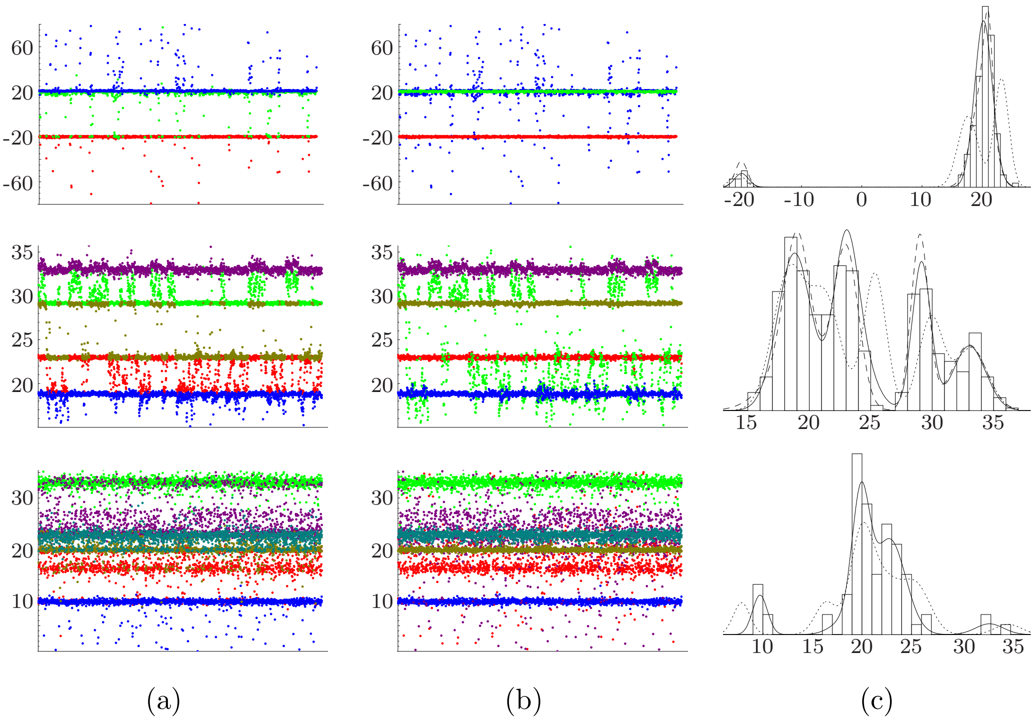

Figure 2: Up: Results for mixture (6); Middle: Results for mixture (7); Down: Results for galaxy dataset. Reordered values according to (a) Pivotal Reordering and (b) ECR algorithms. (c) Plug-in density estimates after reordering according to Pivotal Reordering algorithm (dotted line) and ECR algorithm (solid line) and the true pdf (dashed line).

ergodic averages are presented in the first two rows of Table 1. Clearly, there are major differences between the results obtained by the two reordering schemes and this is due to the fact that for both datasets the posterior distribution has minor modes, i.e., it exhibits genuine multimodality.

Notice that in mixture (6), the second and third components are close to each other. Therefore, the sampler is expected to often combine them to one, leaving one component empty with its parameter values generated from the prior distribution. Indeed, in 20000 iterations (after burn-in), the relative frequency of the existence of an empty component was almost 17%. Since the means' prior variance is large (recall that Richardson and Green's choice for the prior variance is the square of the data midrange), a value for the mean generated from the prior has 95% probability to lie in the interval $(-110, 151)$. So, the generated value for the mean of an empty component may be quite far from the corresponding high posterior density area. Under the standard Pivotal Reordering algorithm, if the generated value from the prior is too small (resp., large) then the empty component will be relabelled as the one corresponding to the smallest (resp., largest)

	Pivotal Reordering						ECR					
$\widehat{\mathbf{E}}(\boldsymbol{\mu}\|\boldsymbol{x})$	23.08	17.62	-20.59				20.46	19.61	-19.96			
	(.419)	(.391)	(.122)				(.016)	(.194)	(.003)			
$\widehat{\mathbf{E}}(\boldsymbol{\sigma}^2\|\boldsymbol{x})$	1.83	2.27	1.42				2.01	2.15	1.36			
	(.278)	(.270)	(.021)				(.025)	(.071)	(.009)			
$\widehat{\mathbf{E}}(\boldsymbol{p}\|\boldsymbol{x})$	.522	.418	.060				.723	.216	.061			
	(.016)	(.016)	(.000)				(.011)	(.011)	(.000)			
$\widehat{\mathbf{E}}(\boldsymbol{\mu}\|\boldsymbol{x})$	21.61	30.06	18.40	33.25	25.82		22.96	25.28	18.80	32.95	29.13	
	(.260)	(.179)	(.101)	(.059)	(.536)		(.005)	(1.007)	(.009)	(.025)	(.007)	
$\widehat{\mathbf{E}}(\boldsymbol{\sigma}^2\|\boldsymbol{x})$	1.47	1.05	1.86	1.69	0.80		1.01	1.57	1.92	1.80	0.56	
	(.115)	(.109)	(.063)	(.064)	(.062)		(.006)	(.056)	(.015)	.(.038)	(.008)	
$\widehat{\mathbf{E}}(\boldsymbol{p}\|\boldsymbol{x})$	.222	.141	.317	.128	.192		.255	.070	.360	.138	.178	
	(.001)	(.010)	(.015)	(.005)	(.007)		(.003)	(.005)	(.007)	.(.003)	(.003)	
$\widehat{\mathbf{E}}(\boldsymbol{\mu}\|\boldsymbol{x})$	7.92	16.35	19.86	22.21	25.53	34.60	9.71	18.29	19.88	22.75	23.00	32.84
	(.100)	(.083)	(.046)	(.038)	(.071)	(.094)	(.002)	(.129)	(.009)	(.015)	(.133)	(.039)
$\widehat{\mathbf{E}}(\boldsymbol{\sigma}^2\|\boldsymbol{x})$	0.70	1.19	1.40	3.17	1.94	1.87	0.57	2.15	0.79	2.63	2.10	2.05
	(.017)	(.034)	(.133)	(.219)	(.068)	(.087)	(.009)	(.101)	(.028)	(.055)	(.115)	(.125)
$\widehat{\mathbf{E}}(\boldsymbol{p}\|\boldsymbol{x})$	.081	.104	.286	.307	.179	.043	.090	.064	.335	.387	.077	.047
	(.000)	(.004)	(.004)	(.004)	(.005)	(.000)	(.000)	(.003)	(.003)	.(.005)	(.003)	(.000)

Table 1: Ergodic averages and their standard errors for the datasets modelled with univariate normal mixtures. Up: mixture (6). Middle: mixture (7). Down: Galaxy dataset.

mean. This clearly leads to underestimation of the smallest and overestimation of the largest mean. Furthermore, in the case where the value generated from the prior lies in the interval $(-20, 20)$, the empty component will be often relabelled as the one corresponding to the middle mean resulting in its underestimation as well. On the contrary, the ECR algorithm explicitly takes care of the above situation. As we can see in Figure 2(b), the resulting reordering succesfully solves the label switching problem by taking into account the minor mode corresponding to the existence of an empty component. Moreover, this produces a better fit since the first two components have their means in the high posterior probability area. The generated values from the prior are always assigned to the third label and so, the extreme values counterbalance each other.

In the case of mixture (7) the large sample size leaves no room for empty components to appear; here, the relative frequency of empty components was only 1.5% and so, the values generated from the prior did not affect much the estimates as in the previous example. Therefore, one would expect the two algorithms to perform similarly. However, this is not the case as we are facing another occasion of genuine multimodality. The reorderings of the means arising by the two algorithms are illustrated in the middle row of Figure 2.

Look first at the means reordering produced by the ECR algorithm. We can see that there are four "stable" components and one with its values gathered in two regions: one around 19, which is the true value of the second component's mean, and one in a seemingly nonsense area in the interval $(30, 33)$. This happens because for many iterations the sampler combined the first two components to one and, instead of creating one empty component, it split the fifth component into two having similar means and different variances. So, we can conclude that there are two competing models with five components that fit well to the data: one with the two first and another with the two last components having nearby (or possibly equal) means. On the other hand, the standard Pivotal Reordering algorithm treats the simulated output as before, and so, the relabelling results in a reordered output that does not highlight the two isolated modes of the posterior distribution at all.

4.3 Galaxy dataset

In this section we demonstrate the performance of our method on the well-known galaxy dataset. The data consist of $n = 82$ galaxy velocities (in 10^3 Km/Sec) diverging from our own, sampled from the conic sections of Corona Borealis. According to Richardson and Green (1997) who fit a mixture of normal distributions, the most probable number of components equals six. Considering the same number of components, we ran the standard random beta model for 60000 iterations (after 10000 iterations for burn-in) and then reordered the output via the ECR and the Pivotal Reordering algorithms. The results for the components means as well as the data histogram together with the corresponding plug-in densities are illustrated in the last row of Figure 2. Moreover, the resulting posterior mean estimates are presented in the last row of Table 1. Notice that, similarly to the first example of the previous subsection, the Pivotal Reordering algorithm results in underestimation and overestimation of the smallest and largest mean, respectively. This is a consequence of the fact that the posterior probability of the existence of at least one empty component is considerably large (over 30%). Note that our estimates for the components' means are in agreement with those reported by Jasra et al. (2005), $\widehat{\mathbf{E}}(\boldsymbol{\mu}|\boldsymbol{x}) = (9.71, 19.01, 19.88, 22.71, 22.86, 32.92)$, obtained

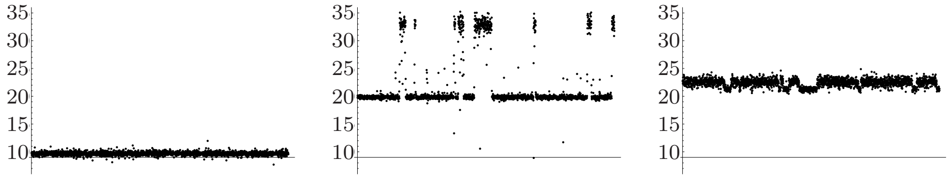

Figure 3: Galaxy dataset: Reordered output for the component means after applying the ECR algorithm for a three t_4 component model.

via the KL divergence based relabelling algorithm of Stephens (1997a, 2000). We have also run Stephens' algorithm for comparison purposes and we found that all parameters' estimates indeed agree. Of course, in this example the truth is unknown but the fact that the two methods give essentially the same answers is clearly favourable for the ECR algorithm since its computational cost is considerably smaller. More specifically, in 20 independent runs of both algorithms the corresponding average CPU times needed for the relabelling part were 1.56 and 256.68 seconds, respectively.

Next, we consider the approach of Stephens (1997a) who modelled the data as a mixture of t_4 distributions and compare our relabelling method with that recently presented by Grün and Leisch (2009). Following them, we fix the number of components to $k = 3$ and ran Stephens' algorithm. The reordered values of component means are plotted in Figure 3. Observe first that the label switching problem is succesfully solved as the reordered simulated values of the means clearly occupy distinct areas. Secondly, the genuine multimodality of the posterior distribution (referred also by Stephens, 1997a, and Grün and Leisch, 2009) is revealed and the high posterior probability areas of the means corresponding to the two modes are succesfully identified. More specifically, we see that the first component mean takes values in a stable region around 9.7 while the other two components switch between 19.8 and 32.8 (second component) and 21.3 and 22.6 (third component). These results are in agreement with those produced by Grün and Leisch (2009).

Grün and Leisch (2009) included a clustering procedure in their algorithm in order to identify the genuine modes of the posterior distribution. We did the same to the reordered output produced by our approach for comparison purposes. More specifically, we applied a K-means clustering algorithm (considering two clusters) to the reordered values of $(\boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ and obtained the results displayed in Table 2. Combining these results with the reordered output in Figure 3, it is obvious that the two genuine posterior modes differ with respect to the second and third components.

Cluster	Weight	μ_1	μ_2	μ_3	σ_1^2	σ_2^2	σ_3^2	p_1	p_2	p_3
1	.831	9.69	19.77	22.56	0.47	0.60	4.05	.093	.319	.588
	(.053)	(.002)	(.035)	(.009)	(.010)	(.029)	(.016)	(.001)	(.002)	(.002)
2	.169	9.70	33.06	21.28	0.78	2.14	3.53	.093	.045	.862
	(.053)	(.003)	(.107)	(.004)	(.027)	(.148)	(.019)	(.000)	(.000)	(.000)

Table 2: Galaxy dataset: Cluster weights and centroids of the reordered MCMC output after applying the ECR algorithm for a three t_4 component model.

Finally, we mention the absolute agreement of the estimated weights of the two clusters with those reported by Grün and Leisch.

4.4 Multivariate normal mixtures

In order to check the perfomance of the proposed method in multivariate settings, we applied the ECR algorithm to MCMC samples generated from the generalization of the random beta model given by Dellaportas and Papageorgiou (2005) considering the number of components to be known. For this purpose, two datasets of bivariate normal mixtures are considered. The first one is a simulated dataset of 200 observations from the distribution $\sum_{j=1}^{4} p_j \mathcal{N}_2(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with actual values shown in Table 3. Notice that this is a challenging case since there are overlapping components. The second one is the version of the Old Faithful dataset analyzed by Stephens (1997a) as well as by Dellaportas and Papageorgiou (2005). The data consist of 272 bivariate observations: the duration of the eruption and the waiting time before the next eruption. According to Dellaportas and Papageorgiou (2005) the most probable number of components equals three.

The scatterplots of the two datasets are shown in Figure 4(c). In the same graph the corresponding plug-in density estimates arising after applying the ECR algorithm to MCMC outputs of size 10000 and 30000, respectively, (after burn-in) are also plotted. Moreover, the reordered values of the component means are shown in Figures 4(a) and (b). As we can see, the samples have been succesfully reordered. The corresponding estimates of the posterior means are in Tables 3 and 4.

For the simulated data the posterior means estimates as produced by the ECR algorithm are quite close to the true values; see Table 3. It is important to note that a constraint on the means would fail to isolate the mode of the posterior as can be concluded from the first row of Figure

parameter	true value	KL	ECR	standard errors
$\boldsymbol{p}$	$(.25, .25, .25, .25)$	$(.30, .21, .24, .25)$	$(.30, .21, .24, .25)$	$(4, 3, 4, 4) \times 10^{-4}$
$\boldsymbol{\mu}$	$(4.5, -2.5)$	$(4.43, -2.36)$	$(4.43, -2.36)$	$(.0014, .0015)$
	$(-3.0, 4.0)$	$(-2.91, 4.04)$	$(-2.91, 4.04)$	$(.0022, .0012)$
	$(6.5, 7.0)$	$(6.73, 7.34)$	$(6.73, 7.34)$	$(.0028, .0036)$
	$(7.0, -3.0)$	$(6.99, -2.77)$	$(6.99, -2.77)$	$(.0036, .0065)$
$\boldsymbol{\Sigma}$	$\begin{pmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.54 & -0.20 \\ -0.20 & 0.81 \end{pmatrix}$	$\begin{pmatrix} 0.54 & -0.20 \\ -0.20 & 0.81 \end{pmatrix}$	$\begin{pmatrix} .0022 & .0013 \\ .0013 & .0027 \end{pmatrix}$
	$\begin{pmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 1.74 & -0.77 \\ -0.77 & 0.69 \end{pmatrix}$	$\begin{pmatrix} 1.74 & -0.77 \\ -0.77 & 0.69 \end{pmatrix}$	$\begin{pmatrix} .0045 & .0026 \\ .0026 & .0018 \end{pmatrix}$
	$\begin{pmatrix} 4 & 2.5 \\ 2.5 & 4 \end{pmatrix}$	$\begin{pmatrix} 3.30 & 2.14 \\ 2.14 & 4.09 \end{pmatrix}$	$\begin{pmatrix} 3.30 & 2.14 \\ 2.14 & 4.09 \end{pmatrix}$	$\begin{pmatrix} .0071 & .0072 \\ .0072 & .0106 \end{pmatrix}$
	$\begin{pmatrix} 4 & 2.5 \\ 2.5 & 9 \end{pmatrix}$	$\begin{pmatrix} 3.55 & 2.27 \\ 2.27 & 10.13 \end{pmatrix}$	$\begin{pmatrix} 3.55 & 2.27 \\ 2.27 & 10.13 \end{pmatrix}$	$\begin{pmatrix} .0097 & .0119 \\ .0119 & .0384 \end{pmatrix}$

Table 3: Posterior means estimates of the parameters for the simulated multivariate dataset according to Stephens' KL algorithm and the ECR algorithm (10000 iterations following a burn-in of 1000).

4(a,b). This is also the case for the variances and covariances (not shown here). Moreover, the high posterior probability area of the weights is close to the area at which they are all equal and thus the components could not be well separated by imposing a constraint on them either. In Table 3 we see that the results completely agree with those obtained by applying the KL relabelling algorithm of Stephens. However, the average CPU time needed by the ECR algorithm for the relabelling part was once more considerably smaller compared to the KL algorithm (0.45 versus 16.31 seconds in 20 independent runs, respectively).

For the Old Faithful dataset the estimates agree with those reported by Dellaportas and Papageorgiou (2005) (see Table 4) who reordered the output by imposing a constraint on the first coordinate of the means. This happens because the simulated values of this coordinate are well separated (see Figure 4). Nevertheless, such artificial IC can be proven quite inefficient in general settings, as discussed previously.

5 Discussion

A simple yet efficient method to solve the label switching problem has been presented. The method uses effectively the natural partition of the allocation space into equivalence classes. Every possible set of the classes representatives $\mathcal{Z}_0$ gives rise to a non-symmetric distribution $f_{\mathcal{Z}_0}$, see (3), that

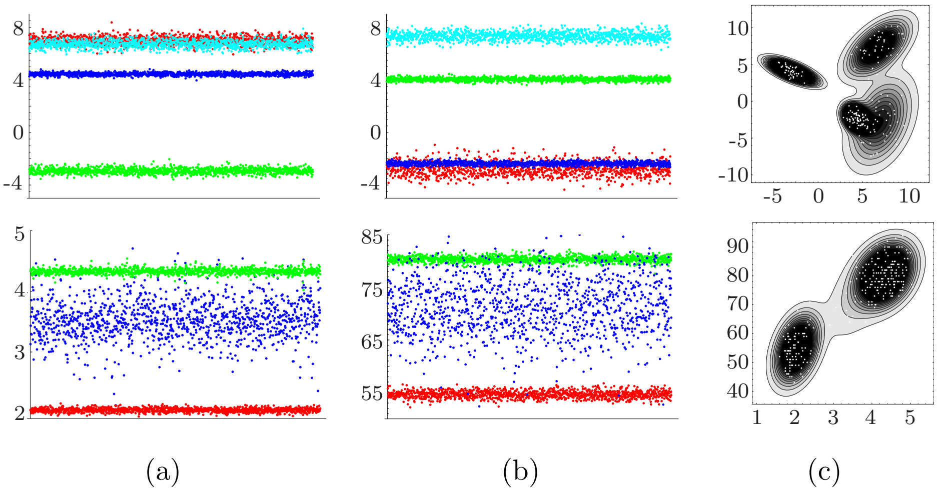

Figure 4: Reordered MCMC outputs of (a) μ_{1j} and (b) μ_{2j}, $j = 1, \ldots, k$, based on ECR algorithm and (c) scatterplot of the bivariate data along with the corresponding plug-in density estimate. Up: Simulated dataset from the mixture in Table 3 ($k = 4$). Down: Old Faithful data ($k = 3$).

can reproduce the posterior distribution. In practice, $\mathcal{Z}_0$ is formed by first selecting a pivotal allocation vector $\boldsymbol{z}^*$ and then minimizing the simple matching distance of each equivalence class from it. In the case where the pivot corresponds to a high probability area of $f(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$, the magnitude of the symmetric modes has totally vanished.

In principle, the determination of $\mathcal{Z}_0$ by the MCMC output itself, seems to be annoying; recall that the convergence stated in Proposition 3.1 occurs for fixed $\mathcal{Z}_0$. Of course, $\mathcal{Z}_0$ could be chosen based on a preliminary run, similarly to what Frühwirth-Schnatter (2001) does in order to select a constraint on the parameter space. But since our approach is based on post-processing the MCMC output, it is clear that reordering a second MCMC sample (according to the selected $\mathcal{Z}_0$) would not make any difference at all.

The proposed reordering method has many desirable properties. First of all, it does not depend on the dimensionality of the parameter space. Secondly, it requires small computational effort compared to other more sophisticated solutions. Third, the distribution of the reordered sample has exactly the same support as the original posterior distribution. This is a very important feature, since it can help to reveal all genuine modes (if any) and does not lead to any serious under– or overestimations of the parameters. Fourth, for all examples we tried, we got essentially the same answers as those reported by the developers of any other “good” approach. Although this can not serve as a formal argument, it is an encouraging fact for the use of the ECR algorithm

parameter	D&P	ECR	standard error
$\boldsymbol{p}$	$(.572, .340, .087)$	$(.590, .351, .059)$	$(.0052, .0004, .0052)$
$\boldsymbol{\mu}$	$(4.34, 80.34)$	$(4.33, 80.33)$	$(.0019, .0129)$
	$(2.02, 54.48)$	$(2.04, 54.63)$	$(.0006, .0080)$
	$(3.44, 70.19)$	$(3.53, 71.33)$	$(.0198, .3182)$
$\boldsymbol{\Sigma}$	$\begin{pmatrix} 0.14 & 0.47 \\ 0.47 & 32.86 \end{pmatrix}$	$\begin{pmatrix} 0.15 & 0.63 \\ 0.63 & 32.46 \end{pmatrix}$	$\begin{pmatrix} .0010 & .0073 \\ .0073 & .1368 \end{pmatrix}$
	$\begin{pmatrix} 0.06 & 0.32 \\ 0.32 & 34.58 \end{pmatrix}$	$\begin{pmatrix} 0.09 & 0.66 \\ 0.66 & 38.55 \end{pmatrix}$	$\begin{pmatrix} .0003 & .0023 \\ .0023 & .0488 \end{pmatrix}$
	$\begin{pmatrix} 0.29 & 3.32 \\ 3.32 & 85.98 \end{pmatrix}$	$\begin{pmatrix} 0.30 & 1.75 \\ 1.75 & 82.07 \end{pmatrix}$	$\begin{pmatrix} .0820 & .7422 \\ .7422 & 16.66 \end{pmatrix}$

Table 4: Old Faithful dataset: Posterior means estimates reported by Dellaportas and Papageorgiou (2005) and based on the ECR algorithm together with their estimated standard errors (30000 iterations following a burn-in of 20000).

since it is by far more simple and less computationally demanding than these approaches.

In all of the examples presented in this paper the number of components k is considered known. However, this does not limit the applicability of the proposed method. Recall that in the case of unknown k where transdimensional MCMC algorithms are used (e.g. the reversible jump MCMC of Richardson and Green, 1997), estimates of the parameters are obtained *conditional* on the number of components, i.e., one set of estimates for each value of k. Similarly, the ECR algorithm must be applied separately to each subset of the output that corresponds to the same k.

In many cases, the original algorithm does not use data augmentation in the first place. For instance, this holds for the Metropolis–Hastings algorithm. However, it is always valid to simulate the allocations *after* having obtained the $(\boldsymbol{p}, \boldsymbol{\theta})^{(i)}$, $i = 1, \ldots, M$, output. Generation of $\boldsymbol{z}^{(i)}$ conditional on $(\boldsymbol{p}, \boldsymbol{\theta})^{(i)}$ (and $\boldsymbol{x}$) under model (1) is straightforward and the augmented sample $(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta})^{(i)}$, $i = 1, \ldots, M$, targets $f(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ as required. Afterwards, the ECR algorithm can be applied to the augmented sample as before. Simulations (not reported here) have shown that everything works as in the previous examples.

It is evident that the ECR algorithm has many common characteristics with previous approaches to the label switching problem. The restriction of the allocation space to $\mathcal{Z}_0$ can be considered analogous to the ICs imposed to the parameter space, see Subsection 2.1. The determination of $\mathcal{Z}_0$ based on a pivot is a modification of the Pivotal Reordering algorithm of Marin et al. (2005). Finally, the fact that the algorithm is applied to the allocation space could be considered as a slight resemblance to Stephens' (2000) KL based approach. The basic difference is that Stephens deals with the similarity of the allocations' estimated posterior distribution rather than the observed allocations themselves. However, the previous approaches are either inefficient or computationally unappealing in practice. For instance, ICs and the default version of the Pivotal Reordering algorithm work well only in cases where the mixture components are far apart. On the other hand, the relabelling algorithms via loss functions and the KL based algorithm of Stephens (2000) are quite elaborate methods, but the high computational cost limits their applicability. Therefore, we strongly suggest ECR algorithm for the solution of the label switching problem since it is both efficient and easy to be applied.

Acknowledgements

The authors wish to thank the two anonymous referees for their helpful comments and suggestions which considerably improved the paper. The first author gratefully acknowledges National Scholarship Foundation of Greece for financial support.

SUPPLEMENTAL MATERIALS

Appendix and Rweave code: The supplemental materials include (a) an appendix with the proofs of Lemma 3.1 and Proposition 3.1 as well as some other technical results and (b) the file `ecr_urb.Rnw` which contains an `Rweave` code that can be used to replicate the analysis for the simulated datasets from the mixtures in Section 4.2 and for the galaxy dataset as well as its companion file `Readme.pdf`.

References

Carpaneto, P. (1980). Algorithm 548: Solution of the assignment problem. *ACM Transactions on Mathematical Software*, **6**, 104–111.

Casella, G., Robert, C.P. and Wells, M. (2004). Mixture models latent variables and partitioned importance sampling. *Statistical Methodology*, **1**, 1–18.

Celeux, G. (1997). Discussion of "On Bayesian analysis of mixtures with an unknown number of components" by S. Richardson and P.J. Green. *Journal of the Royal Statistical Society, Series B*, **59**, 775–776.

Celeux, G., Hurn, M. and Robert, C.P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95**, 957–970.

Dellaportas, P. and Papageorgiou, I. (2005). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing*, **16**, 57–68.

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the American Statistical Association*, **96**, 194–209.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the Royal Statistical Society, Series B*, **56**, 363–375.

Grün, B., and Leisch, F. (2009). Dealing with label switching in mixture models under genuine multimodality. *Journal of Multivariate Analysis*, **100**, 851–861.

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

Hurn, M., Justel, A. and Robert, C.P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, **12**, 55–79.

Marin, J.M., Mengersen, K. and Robert, C.P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, **25**, D. Dey and C.R. Rao (eds). Elsevier-Sciences.

Marin, J.M. and Robert, C.P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer-Verlag, New York.

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

Papaspiliopoulos, O. and Roberts, G.O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95**, 169–186.

Stephens, M. (1997a). Bayesian methods for mixtures of normal distributions. D. Phil dissertation, Dept. Statistics, Univ. Oxford.

Stephens, M. (1997b). Discussion of "On Bayesian analysis of mixtures with an unknown number of components" by S. Richardson and P.J. Green. *Journal of the Royal Statistical Society, Series B*, **59**, 768–769.

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

Appendix to
"An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions"

published in the *Journal of Computational and Graphical Statistics*

Panagiotis Papastamoulis and George Iliopoulos[1]

Lemma A.1. *The posterior distribution of* $(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta})$ *satisfies*

$$f(\tau\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{z}, \tau(\boldsymbol{p}, \boldsymbol{\theta})|\boldsymbol{x}), \quad \forall\tau \in \mathcal{T}_k. \tag{A.1}$$

Moreover, for the conditional distribution of $(\boldsymbol{p}, \boldsymbol{\theta})$ *given* $(\boldsymbol{x}, \boldsymbol{z})$ *it holds*

$$f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \tau\boldsymbol{z}) = f(\tau(\boldsymbol{p}, \boldsymbol{\theta})|\boldsymbol{x}, \boldsymbol{z}), \quad \forall\tau \in \mathcal{T}_k. \tag{A.2}$$

Proof. Let $I = \{1, \ldots, n\}$. For any $\boldsymbol{z} \in \mathcal{Z}$ write $I = I_1(\boldsymbol{z}) \cup \cdots \cup I_k(\boldsymbol{z})$ with $I_j(\boldsymbol{z}) = \{i : z_i = j\}$ and let $n_j(\boldsymbol{z}) = \text{card}\,(I_j(\boldsymbol{z})) = \sum_{i=1}^{n} I(z_i = j)$, $j = 1, \ldots, k$. Also, let $g(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) := \prod_{j=1}^{k} \prod_{i \in I_j(z)} f(x_i|\theta_j) p_j^{n_j(\boldsymbol{z})}$. Then, we can write $f(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{p}) f(\boldsymbol{p}, \boldsymbol{\theta})/f(\boldsymbol{x}) = g(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{p}, \boldsymbol{\theta})/f(\boldsymbol{x})$. Hence, for every $\tau \in \mathcal{T}_k$ we have that

$$f(\tau\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = g(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{p}, \boldsymbol{\theta})/f(\boldsymbol{x}). \tag{A.3}$$

Let $\tau^{-1} = (t'_1, \ldots, t'_k)$ be the reverse permutation of $\tau = (t_1, \ldots, t_k)$. Observe that $I_j(\tau\boldsymbol{z}) = \{i : t_{z_i} = j\} = \{i : z_i = t'_j\} = I_{t'_j}(\boldsymbol{z})$. Hence,

$$\begin{aligned} g(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \tau\boldsymbol{z}) = & \prod_{j=1}^{k} \prod_{i \in I_j(\tau\boldsymbol{z})} f(x_i|\theta_j) p_j^{n_j(\tau\boldsymbol{z})} = \prod_{j=1}^{k} \prod_{i \in I_{t'_j(\boldsymbol{z})}} f(x_i|\theta_j) p_j^{n_{t'_j}(\boldsymbol{z})} \\ = & \prod_{j=1}^{k} \prod_{i \in I_j(\boldsymbol{z})} f(x_i|\theta_{t_j}) p_{t_j}^{n_j(\boldsymbol{z})} = g(\tau(\boldsymbol{p}, \boldsymbol{\theta})|\boldsymbol{x}, \boldsymbol{z}). \end{aligned} \tag{A.4}$$

[1] Corresponding author. Department of Statistics and Insurance Science, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece e-mail: geh@unipi.gr

Now, notice that the prior distribution is invariant with respect to the labelling, that is, $f(\boldsymbol{p},\boldsymbol{\theta}) = f(\tau(\boldsymbol{p},\boldsymbol{\theta}))$, $\forall\tau \in \mathcal{T}_k$. Substituting this together with (A.4) into (A.3) we get $f(\tau\boldsymbol{z},\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{z},\tau(\boldsymbol{p},\boldsymbol{\theta})|\boldsymbol{x})$ and the proof of (A.1) is completed. Finally, (A.2) follows immediately from (A.1) and the fact that $f(\boldsymbol{z},\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x}) = w(\boldsymbol{z}|\boldsymbol{x})f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z})$ and $w(\boldsymbol{z}|\boldsymbol{x})$ is invariant with respect to the permutations of the labels. □

Proof of Lemma 3.1. After rearranging the k^n terms, (2) can be written as

$$f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{\boldsymbol{z}\in\mathcal{Z}_0}\sum_{\boldsymbol{z}^*\in\Xi_{\boldsymbol{z}}} w(\boldsymbol{z}^*|\boldsymbol{x})f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z}^*). \tag{A.5}$$

But

$$\begin{aligned}\sum_{\boldsymbol{z}^*\in\Xi_{\boldsymbol{z}}} w(\boldsymbol{z}^*|\boldsymbol{x})f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z}^*) &= \sum_{\tau\in\mathcal{T}_k}\frac{w(\tau\boldsymbol{z}|\boldsymbol{x})}{(k-k_0(\tau\boldsymbol{z}))!}f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x},\tau\boldsymbol{z})\\ &= \frac{w(\boldsymbol{z}|\boldsymbol{x})}{(k-k_0(\boldsymbol{z}))!}\sum_{\tau\in\mathcal{T}_k} f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x},\tau\boldsymbol{z}). \qquad (A.6)\end{aligned}$$

Substituting (A.2) and (A.6) into (A.5) we get

$$\begin{aligned}f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x}) =& \sum_{\boldsymbol{z}\in\mathcal{Z}_0}\frac{w(\boldsymbol{z}|\boldsymbol{x})}{(k-k_0(\boldsymbol{z}))!}\sum_{\tau\in\mathcal{T}_k} f(\tau(\boldsymbol{p},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{z})\\ =& \frac{1}{k!}\sum_{\tau\in\mathcal{T}_k} k!\sum_{\boldsymbol{z}\in\mathcal{Z}_0}\frac{w(\boldsymbol{z}|\boldsymbol{x})}{(k-k_0(\boldsymbol{z}))!}f(\tau(\boldsymbol{p},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{z}) = \frac{1}{k!}\sum_{\tau\in\mathcal{T}_k} f_{\mathcal{Z}_0}(\tau(\boldsymbol{p},\boldsymbol{\theta})|\boldsymbol{x})\end{aligned}$$

as stated, and this completes the proof. □

Lemma A.2. *Let $(\boldsymbol{z},\boldsymbol{p},\boldsymbol{\theta}) \sim f(\boldsymbol{z},\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x})$ and, conditional on $(\boldsymbol{z},\boldsymbol{p},\boldsymbol{\theta})$, τ has the uniform distribution on $\mathcal{T}_{\boldsymbol{z}}$ defined in Section 3. Then, $\tau^{-1}(\boldsymbol{p},\boldsymbol{\theta}) \sim f_{\mathcal{Z}_0}(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x})$.*

Proof. Observe first that τ depends solely on $\boldsymbol{z}$ and there are exactly $(k-k_0(\boldsymbol{z}))!$ permutations that switch $\boldsymbol{z}$ to $\mathcal{Z}_0$. Thus, the joint pdf of $\tau,\boldsymbol{z},\boldsymbol{p},\boldsymbol{\theta}$ is

$$f(\tau,\boldsymbol{z},\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z})f(\boldsymbol{z}|\boldsymbol{x})f(\tau|\boldsymbol{z}) = \frac{w(\boldsymbol{z}|\boldsymbol{x})}{(k-k_0(\boldsymbol{z}))!}f(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z})I(\tau\in\mathcal{T}_{\boldsymbol{z}}). \tag{A.7}$$

Let $(\tau^*, \boldsymbol{z}^*, \boldsymbol{p}^*, \boldsymbol{\theta}^*) = (\tau, \tau\boldsymbol{z}, \tau^{-1}(\boldsymbol{p}, \boldsymbol{\theta}))$. Then, for any $\tau \in \mathcal{T}_k, \boldsymbol{u} \in \mathcal{Z}_0$ and measurable subset C of the space $\mathcal{A}$ (say) of $(\boldsymbol{p}, \boldsymbol{\theta})$ we have

$$\mathbf{P}(\tau^* t, \boldsymbol{z}^* = \boldsymbol{u}, (\boldsymbol{p}^*, \boldsymbol{\theta}^*) \in C|\boldsymbol{x}) = \mathbf{P}(\tau = t, \tau\boldsymbol{z} = \boldsymbol{u}, \tau^{-1}(\boldsymbol{p}, \boldsymbol{\theta}) \in C|\boldsymbol{x}) =$$
$$\mathbf{P}(\tau = t, \boldsymbol{z} = \tau^{-1}\boldsymbol{u}, (\boldsymbol{p}, \boldsymbol{\theta}) \in \tau C|\boldsymbol{x}) = \frac{w(t^{-1}\boldsymbol{u}|\boldsymbol{x})}{(k - k_0(t^{-1}\boldsymbol{u}))!} \int_{tC} f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, t^{-1}\boldsymbol{u}) \mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{\theta},$$

since $t \in T_{t^{-1}\boldsymbol{u}}$ for all $t \in \mathcal{T}_k$ and $\boldsymbol{u} \in \mathcal{Z}_0$. But $w(t^{-1}\boldsymbol{u}|\boldsymbol{x}) = w(\boldsymbol{u}|\boldsymbol{x})$ and $k_0(t^{-1}\boldsymbol{u}) = k_0(\boldsymbol{u})$, so, using also (A.2), the last expression becomes

$$\frac{w(\boldsymbol{u}|\boldsymbol{x})}{(k - k_0(\boldsymbol{u}))!} \int_{tC} f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, t^{-1}\boldsymbol{u}) \mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{\theta} = \frac{w(\boldsymbol{u}|\boldsymbol{x})}{(k - k_0(\boldsymbol{u}))!} \int_{C} f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{u}) \mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{\theta}.$$

Hence, the density of $(\tau^*, \boldsymbol{z}^*, \boldsymbol{p}^*, \boldsymbol{\theta}^*)$ (with respect to the appropriate product measure) is

$$f^*(t, \boldsymbol{u}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = \frac{w(\boldsymbol{u}|\boldsymbol{x})}{(k - k_0(\boldsymbol{u}))!} f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{u}), \quad t \in \mathcal{T}_k, \boldsymbol{u} \in \mathcal{Z}_0, (\boldsymbol{p}, \boldsymbol{\theta}) \in \mathcal{A}, \quad \text{(A.8)}$$

while the marginal distribution of $\boldsymbol{p}^*, \boldsymbol{\theta}^* = \tau^{-1}(\boldsymbol{p}, \boldsymbol{\theta})$ is

$$f^*(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}) = \sum_{t \in \mathcal{T}_k} \sum_{\boldsymbol{u} \in \mathcal{Z}_0} \frac{w(\boldsymbol{u}|\boldsymbol{x})}{(k - k_0(\boldsymbol{u}))!} f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{u}) = k! \sum_{\boldsymbol{u} \in \mathcal{Z}_0} \frac{w(\boldsymbol{u}|\boldsymbol{x})}{(k - k_0(\boldsymbol{u}))!} f(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{u})$$

i.e., $f_{\mathcal{Z}_0}(\boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ as stated. □

Proof of Proposition 3.1. Clearly, the augmented sequence $(\tau_i, \boldsymbol{z}^{(i)}, (\boldsymbol{p}, \boldsymbol{\theta})^{(i)})$ is a Markov chain with limit distribution $f(\tau, \boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\theta}|\boldsymbol{x})$ in (A.7). Now, $(\tau^*, \boldsymbol{z}^*, \boldsymbol{p}^*, \boldsymbol{\theta}^*)^{(i)}$ is an invertible transformation of $(\tau_i, \boldsymbol{z}^{(i)}, (\boldsymbol{p}, \boldsymbol{\theta})^{(i)})$, so the corresponding sequence is a Markov chain as well with limit distribution $f^*(\tau^*, \boldsymbol{z}^*, \boldsymbol{p}^*, \boldsymbol{\theta}^*|\boldsymbol{x})$ in (A.8). The result follows again after marginalization. □