Sampling from Multimodal Posteriors: An Overview

Ezequiel B. Santos[*]

School of Applied Mathematics - Getulio Vargas Foundation (FGV EMAp)

Abstract

In recent years, Bayesian models have gained significant popularity for their ability to incorporate prior knowledge and provide probabilistic interpretations of uncertainty. This is achieved through the use of Markov Chain Monte Carlo (MCMC) methods, which have been improved and adapted to handle complex posterior distributions. However, one of the main challenges in Bayesian inference arises from the *label switching* problem, particularly in mixture models where the posterior distribution is multimodal and non-identifiable. This paper provides an overview of the label switching problem, its implications for MCMC sampling, and various strategies to address it. We focus on three MCMC methods: simulated tempering, parallel tempering, and tempered transitions. We illustrate the two last methods through a simulated data from a random beta model, comparing the effectiveness of each method in sampling from the posterior distribution. We also discuss the impact of label switching on posterior inference and the effectiveness of post-processing techniques to resolve it. Our findings highlight the strengths and limitations of each method, providing insights into their practical applications in Bayesian inference.

Keywords: Bayesian inference, Markov Chain Monte Carlo, label switching, multimodal posteriors, mixture models, simulated tempering, parallel tempering, tempered transitions.

1 Introduction

Bayesian mixture models are widely used for modeling heterogeneous data where observations are assumed to arise from multiple latent subpopulations. A common example is the finite mixture of Gaussian distributions, which can flexibly approximate complex, multimodal densities. Despite their practical utility, these models pose significant computational challenges, particularly in posterior inference via Markov Chain Monte Carlo (MCMC) methods.

One of the main obstacles is the *label switching* problem [Stephens, 2002], which arises due to the non-identifiability of the mixture components. The likelihood and posterior distributions are invariant to permutations of component labels, resulting in a posterior with multiple symmetric modes—one for each of the $K!$ permutations of the component labels. Standard MCMC algorithms, such as Gibbs sampling, often struggle to move between these modes, leading to poor mixing and biased inference.

To address this issue, several strategies have been proposed. One class of solutions involves modifying the sampling algorithm to better explore multimodal posteriors. Techniques such as *simulated tempering* [Geyer and Thompson, 1995], *parallel tempering* [Earl and Deem, 2005], and *tempered transitions* [Neal, 1996] have been developed to improve the mixing of MCMC chains in the presence of multiple modes. These methods introduce auxiliary distributions at different "temperatures" to allow the sampler to traverse the posterior landscape more effectively.

[*] ezequiel.braga.santos@gmail.com

Another class of solutions aims to resolve the label switching problem through post-processing, either by imposing identifiability constraints or by applying relabeling algorithms that align the sampled parameters across iterations [Jasra et al., 2005].

In this work, we investigate the performance of different tempering-based MCMC methods for sampling from the posterior of a Gaussian mixture model. We simulate data from a known mixture distribution and compare the effectiveness of standard Gibbs sampling, parallel tempering, and tempered transitions. We also explore post-processing approaches for resolving label switching and assess their impact on the posterior summaries. Our goal is to provide a comprehensive comparison of these methods in terms of their ability to explore multimodal posteriors, their computational efficiency, and the quality of their inference.

2 Label Switching

As defined by Jasra et al. [Jasra et al., 2005], let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be a random sample from a mixture model with K components:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(x_i \mid \phi_k),$$

where π_k are the mixing proportions, and $f(x_i \mid \phi_k)$ are the component densities with parameters ϕ_k. We assume $\sum_{k=1}^{K} \pi_k = 1$ and $0 \leq \pi_k \leq 1$ for all k.

Let $\boldsymbol{\theta} = ((\pi_1, \phi_1), \ldots, (\pi_K, \phi_K))$, and let $\sigma \in S_K$ be a permutation of $\{1, \ldots, K\}$. Then,

$$\sigma(\boldsymbol{\theta}) = \left(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(K)}\right).$$

The *label switching* problem arises because the likelihood and posterior are invariant to such permutations. That is,

$$p(\boldsymbol{x} \mid \sigma(\boldsymbol{\theta})) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{\sigma(k)} f(x_i \mid \phi_{\sigma(k)}),$$

is equal to $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ for all $\sigma \in S_K$. This symmetry introduces $K!$ identical modes in the posterior distribution.

2.1 Example: Random Beta Model

A classic example is the random beta model, where we consider a mixture of K Normal components

$$x_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma} \sim \sum_{k=1}^{K} \pi_k \,\text{Normal}(x_i \mid \mu_k, \sigma_k^2),$$

with prior distributions

$$\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K), \\
\mu_k &\sim \text{Normal}(m, \tau^2), \\
\sigma_k^2 &\sim \text{InverseGamma}(a, b),
\end{aligned}$$

for $k = 1, \ldots, K$. These distributions lead to a semi-conjugate prior structure and a posterior that is multimodal due to label switching.

3 MCMC for Multimodal Posteriors

The multimodality caused by label switching creates difficulties for Markov Chain Monte Carlo (MCMC) methods. Standard algorithms may get trapped in local modes, resulting in poor mixing and high autocorrelation. The most common approach is to use Gibbs sampling, which consists to introduce latent variables $\boldsymbol{z} = (z_1, \ldots, z_n)$ indicating the component assignments:

$$\mathbb{P}(z_i = k \mid \boldsymbol{\pi}) = \pi_k.$$

Then we can compute the full conditional posterior distributions of $(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\phi})$ and iterate over them. However, the sampler may get stuck in one mode and fail to explore others, making it ineffective for label-switching scenarios. Below, we explore how various MCMC techniques handle these challenges and illustrate their effectiveness in sampling from a simulated data.

3.1 Simulated Tempering

Following the idea proposed by Geyer [Geyer and Thompson, 1995], suppose we want to sample from a distribution P with density $p_0(x)$ which is multimodal. We can introduce a sequence of inverse temperatures β_i to create a family of tempered distributions:

$$p_i(x) = \frac{p_0(x)^{\beta_i}}{Z_i},$$

where $\beta_0 = 1 > \beta_1 > \cdots > \beta_m = \beta^*$ and $Z_i = Z_i(\beta_i) = \int p_0(x)^{\beta_i} dx$ is the normalization constant.

The algorithm alternates between:

1. Sampling x from $p_i(x)$ using MCMC,

2. Proposing a change to temperature $j = i \pm 1$ using transition probabilities q_{ij},

3. Accepting the move with probability:

$$\alpha_{ij} = \min\left(1, \frac{p_j(x)\pi_0(j)q_{ji}}{p_i(x)\pi_0(i)q_{ij}}\right),$$

where $\pi_0(i)$ is a pseudoprior over the temperatures. When q_{ij} is uniform, i.e., $q_{ij} = 1/2$ for $1 < i < m$ and $q_{12} = q_{m,m-1} = 1$, we can simplify the acceptance probability.

This method updates the chain in two steps: first, it samples from the tempered distribution $p_i(x)$, and then it proposes to change the temperature. Then, the invariant distribution of the chain for a fixed temperature i keeps $p_i(x)$. So a sample from $p_0(x)$ can be obtained by running the chain for a long time and discarding the samples from $i \neq 0$.

As also noted by Neal [Neal, 1996], a common choice is $\pi_0(i) = 1/Z_i$ to assign uniform prior probabilities to the temperatures. However, in practice, Z_i is often unknown, and need to be estimated. This leads to the need for calibration techniques to ensure that the temperatures are chosen appropriately to balance exploration and convergence. So the algorithm might require a large computational cost.

4 Parallel Tempering

Alternatively, we can run multiple chains in parallel at different temperatures, known as *parallel tempering* [Earl and Deem, 2005]. This method consider the same family of tempered as above, but instead of simulating a single chain, we run m chains at temperatures $\beta_1, \ldots, \beta_m$. Each chain samples from the tempered distribution $p_i(x)$. But, instead of accepting moves between temperatures, we periodically propose to swap states of two chains (i, j): we sample a pair of chains and propose to swap their states with acceptance probability:

$$\alpha_{ij} = \min\left(1, \frac{p_j(x_i)p_i(x_j)}{p_i(x_i)p_j(x_j)}\right).$$

This procedure does not change the invariant distribution of the chain, as it still samples from the tempered distributions $p_i(x)$. In particular, the chain at temperature $i = 0$ will converge to the target distribution $p_0(x)$.

5 Tempered Transitions

Another approach to deal with multimodal posteriors is to use *tempered transitions* [Neal, 1996]. The idea is to construct kernels $\hat{T}_i$ and $\check{T}_i$ that simulate forward and backward transitions, respectively, between the states of a Markov chain at temperatures i and $m-i+1$. The forward kernel $\hat{T}_i$ samples from the tempered distribution $p_i(x)$, while the backward kernel $\check{T}_i$ samples from $p_{m-i+1}(x)$. They must satisfy the detailed balance condition:

$$\hat{T}_i(x, y)p_i(x) = \check{T}_i(y, x)p_{m-i+1}(y),$$

where x, y are states of the Markov chain. The algorithm proceeds as follows:

1. From x_t, simulate $\hat{x}_1, \ldots, \hat{x}_m$ via forward kernels $\hat{T}_i$;

2. Then simulate $\check{x}_m, \ldots, \check{x}_1, \hat{x}_0$ using backward kernels $\check{T}_i$;

3. Accept the proposal $\hat{x}_0$ with probability:

$$\alpha = \min\left(1, \frac{\prod_{i=1}^m p_i(\hat{x}_{i-1}) \prod_{i=1}^m p_{m-i+1}(\check{x}_i)}{\prod_{i=1}^m p_i(\check{x}_{i-1}) \prod_{i=1}^m p_{m-i+1}(\hat{x}_i)}\right).$$

This leads to a Markov chain with invariant distribution $p_0(x)$, as shown by Neal [Neal, 1996].

6 Dealing with Label Switching

While tempering methods improve exploration, they do not resolve label switching. The posterior remains multimodal and non-identifiable. Solutions involve either imposing identifiability constraints or using post-processing.

6.1 Imposing Constraints

A straightforward solution is to enforce parameter ordering, e.g., $\mu_1 < \mu_2 < \cdots < \mu_K$, as in Stephens et al. [Stephens, 1997]. After sampling, parameters are reordered to match this constraint. However, as discussed by Sperring et al. [Sperrin et al., 2010], although this approach works well in many cases and leads to correct marginal posterior distributions, it can lead to biased estimates.

6.2 Relabeling Algorithms

A more general approach uses relabeling algorithms that minimize an expected loss. Let $L_0(a, \sigma(\boldsymbol{\theta}))$ be a base loss function comparing action a with permuted parameter $\sigma(\boldsymbol{\theta})$. Define the symmetrized loss:

$$L(a, \boldsymbol{\theta}) = \min_{\sigma \in S_K} L_0(a, \sigma(\boldsymbol{\theta})).$$

Then, the optimal action a^* solves:

$$a^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{x})}[L(a, \boldsymbol{\theta})].$$

As this expectation is intractable, it is approximated using Monte Carlo methods over MCMC samples. As described by Stephens et al. [Stephens, 2002], the algorithm proceeds by starting with an initial guess for the labels, then iteratively minimizes the posterior expected loss until convergence.

7 Simulated Data

We will simulate data from a random beta model with $K = 4$ components, using the following parameters:

$$\begin{aligned}
\pi_k &= 0.25, \quad k = 1, \ldots, 4, \\
\boldsymbol{\mu} &= (-3, 0, 3, 6), \\
\sigma_k^2 &= 0.55^2, \quad k = 1, \ldots, 4.
\end{aligned}$$

We will compare the performance of standard Gibbs sampling with tempered transitions and parallel tempering. The goal is to analyze how well each method explores the multimodal posterior and handles label switching. We will also visualize the results using trace plots and posterior distributions.

We run all methods for $2 \cdot 10^5$ iterations, discarding the first $5 \cdot 10^4$ as burn-in. Also, we apply tempered distributions for each conditional posterior, i.e.,

$$\begin{aligned}
p(\boldsymbol{\mu} \mid \boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2) &\propto p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2) p(\boldsymbol{\mu}), \\
p(\boldsymbol{\sigma}^2 \mid \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\pi}) &\propto p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2) p(\boldsymbol{\sigma}^2), \\
p(\boldsymbol{\pi} \mid \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) &\propto p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2) p(\boldsymbol{\pi}),
\end{aligned}$$

because the use of the joint posterior may lead to small acceptance rates, as noted by Jasra et al. [Jasra et al., 2005]. Another important point is that we reparameterize the weights $\boldsymbol{\pi}$ using $\pi_k = \nu_k / \sum_{j=1}^{K} \nu_j$, where $\nu_k \sim \text{Gamma}(1, 1)$, to avoid issues with proposed values outside the simplex.

For the proposal distributions, we use a Gaussian random walk for $\boldsymbol{\mu}$ and reflective random walk for $\boldsymbol{\sigma}$ and $\boldsymbol{\pi}$. The temperature ladder is chosen as $\beta_i = 2^{-i}$ for $i = 0, 1, \ldots, 19$, for tempered transitions, and $\beta_0 = 1, \beta_1 = 0.9, \ldots, \beta_9 = 0.1$ for parallel tempering. Futhermore, we propose a swap move or run the tempered transitions with probability 0.5 every iteration. For full details on the implementation, see the code repository `https://github.com/EzequielEBS/multimodal-bayesian-sampling`.

In the figure 1, we show the simulated data distribution, which is a mixture of four Normal components. The modes are clearly separated, and the data is generated from a well-defined multimodal distribution.

Also, we show the trace plots for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, and $\boldsymbol{\pi}$ obtained from the three methods in the figures 2, 3, and 4.

Next, we show the posterior distributions for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, and $\boldsymbol{\pi}$ after applying the identifiability constraints to the samples obtained from the three methods in the figures 5, 6, and 7.

Lastly, we show the acceptance rates for the three methods in the table 1 and the effective sample sizes in the table 2.

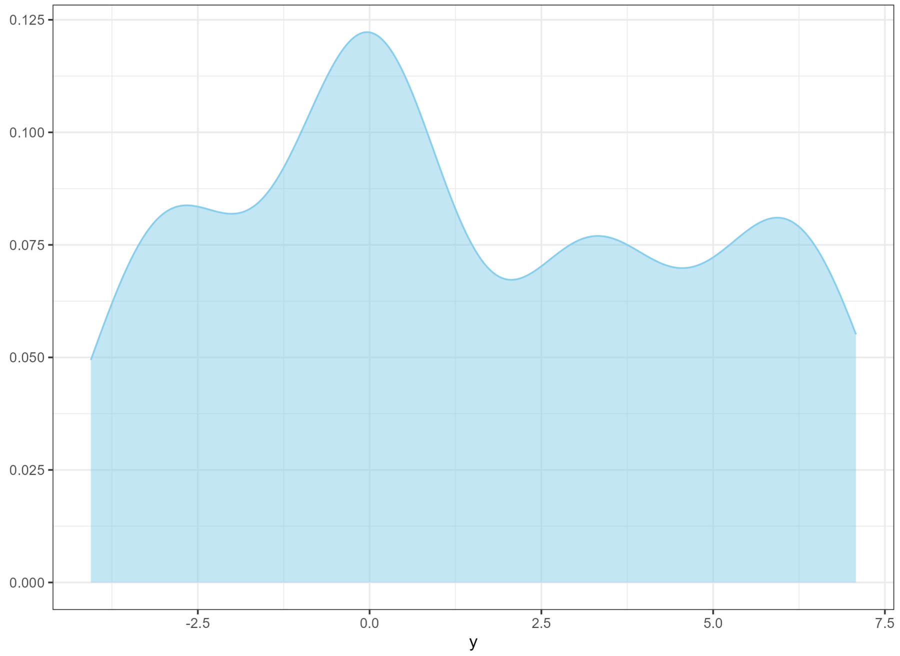

Figure 1: **Simulated Data Distribution.**

Table 1: **Acceptance rates by method.**

Method	μ	σ^2	π
Gibbs	1.000	1.000	1.000
Parallel Tempering	0.265	0.629	0.258
Temperated Transition	0.215	0.377	0.362

Table 2: **Effective sample sizes by method.**

Method	Gibbs	Parallel Tempering	Temperated Transition
μ_1	129953	1005	9
μ_2	124791	1046	10
μ_3	100228	968	5239
μ_4	135154	1013	121
σ_1^2	101667	2942	2289
σ_2^2	88399	3326	141
σ_3^2	72239	3630	3862
σ_4^2	106736	3168	80
π_1	144709	1249	642
π_2	138560	1346	721
π_3	139232	1236	11194
π_4	146301	1275	1070

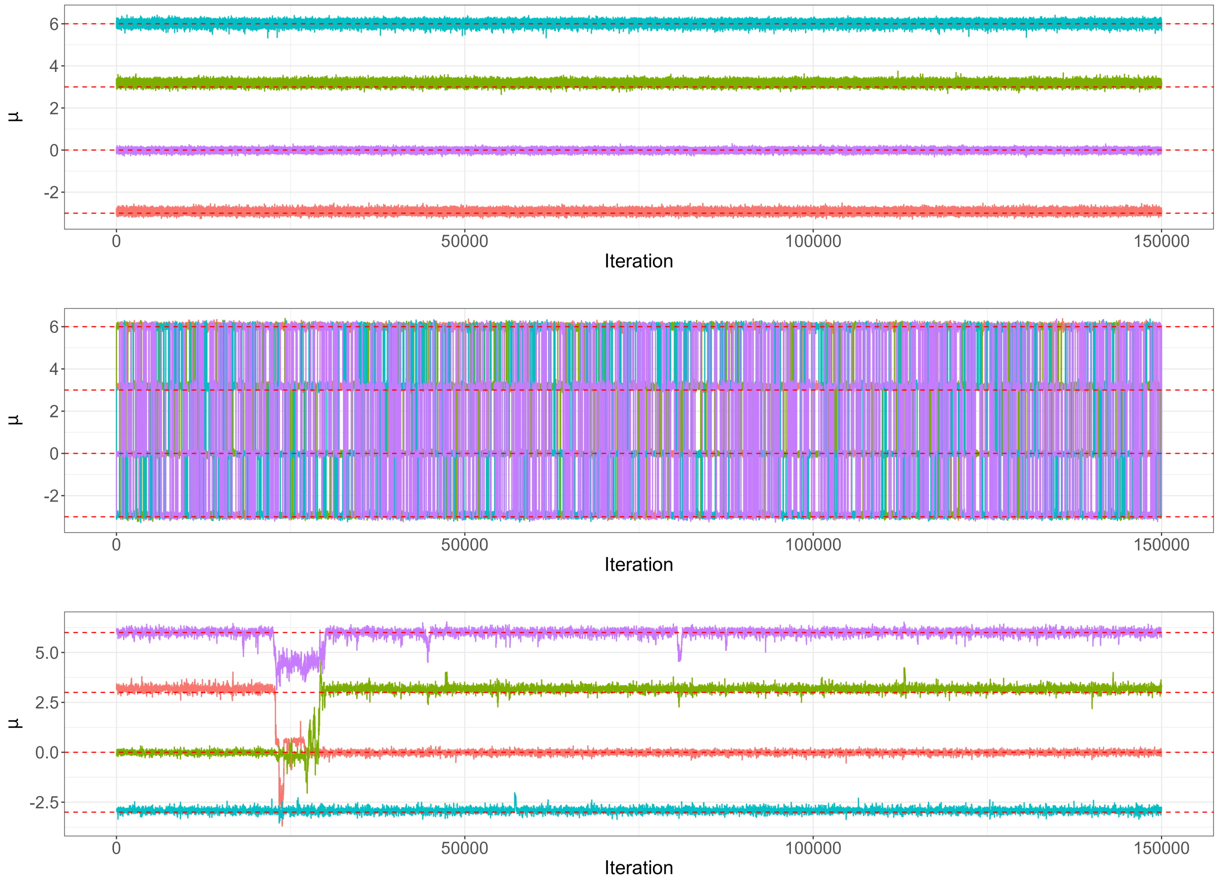

Figure 2: **Trace plots for $\boldsymbol{\mu}$.** The first row shows the trace for Gibbs sampling, the second row for parallel tempering, and the third row for tempered transitions. Each color represents a different component. The dashed lines indicate the true parameter values.

8 Discussion and Conclusion

In this paper, we explored the challenges of sampling from multimodal posteriors in Bayesian inference, focusing on the label switching problem. We discussed how standard MCMC methods, such as Gibbs sampling, struggle with local modes and poor mixing in this context. We then introduced several advanced techniques, including simulated tempering, parallel tempering, and tempered transitions, which improve exploration of the posterior landscape. Also, we discussed how these methods can be combined with relabeling algorithms to address the label switching problem. Then we presented an example using simulated data from a random beta model, demonstrating the effectiveness of these methods in sampling from a multimodal posterior.

We found that while Gibbs sampling does not move between modes, both parallel tempering and tempered transitions are able to explore the multimodal posterior effectively, especially for the parallel tempering method. We also observed that the tempered transitions method has a limited movement between modes, maybe due to the choice of temperatures and the proposal distributions. The acceptance rates were significantly good for all methods, but the effective sample sizes were much lower for the tempered transitions method, indicating that there is a large autocorrelation in the samples and that the method does not explore the posterior well.

Despite the issues related, all methods were able to recover the true parameter values except for some components of $\boldsymbol{\pi}$. Furthermore, the identifiability constraints were effective in aligning the modes of the posterior distributions in this case study. However, the larger computational cost of the tempered methods and the need for a large number of iterations to achieve conver-

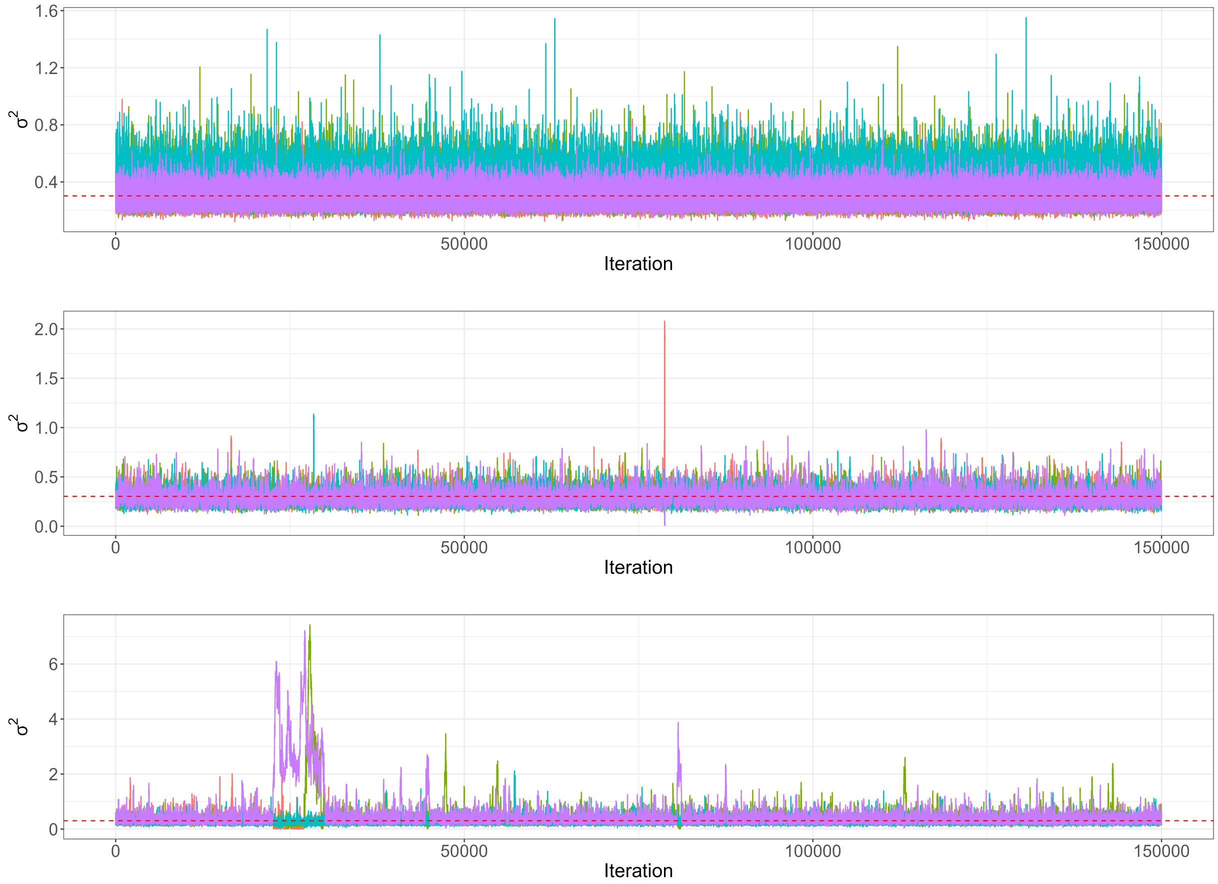

Figure 3: **Trace plots for $\boldsymbol{\sigma^2}$.** The first row shows the trace for Gibbs sampling, the second row for parallel tempering, and the third row for tempered transitions. Each color represents a different component. The dashed lines indicate the true parameter values.

gence are important considerations in practice. Also, we note that the choice of temperatures and proposal distributions is crucial for the performance of tempered methods. In practice, it is often necessary to calibrate these parameters to ensure that the sampler explores the posterior effectively. Future work could explore adaptive methods for selecting temperatures and proposal distributions, as well as more sophisticated relabeling algorithms that can handle more complex multimodal posteriors.

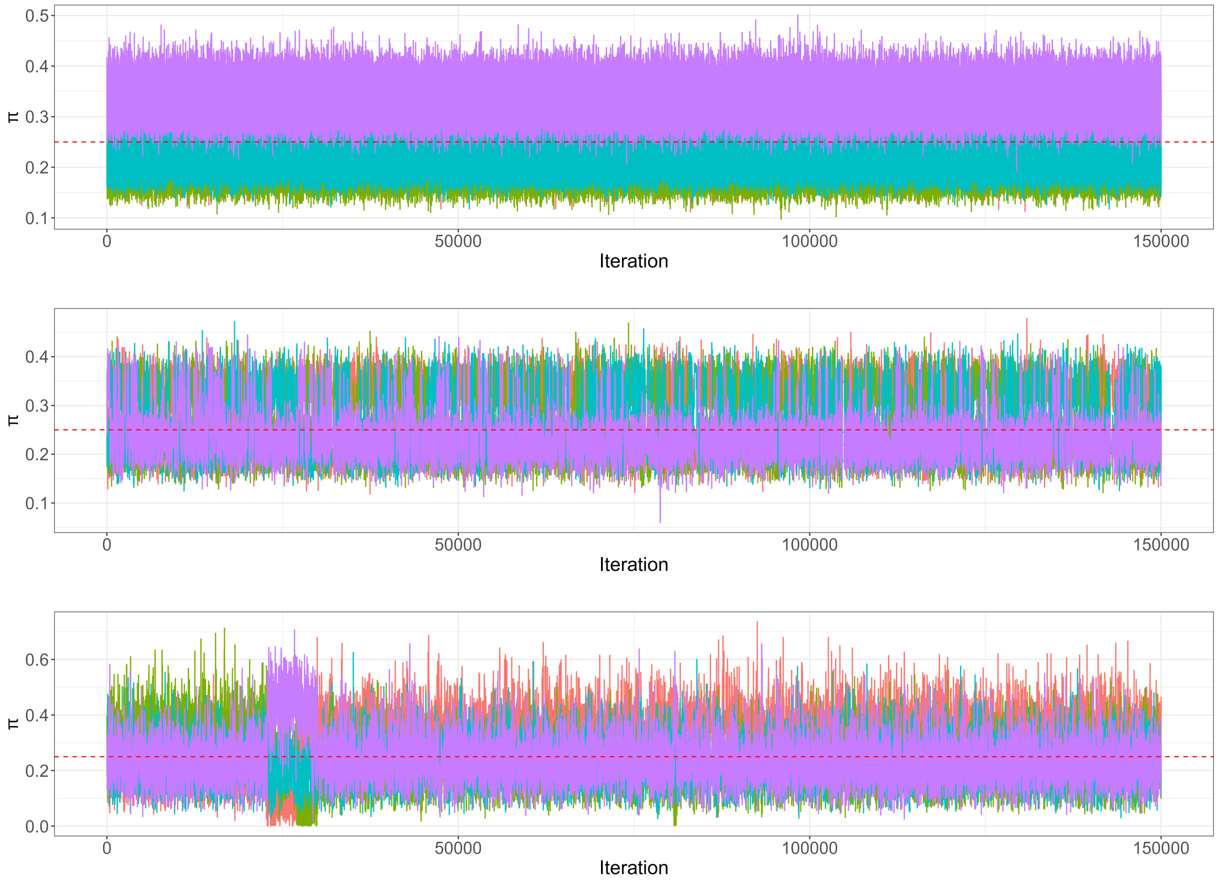

Figure 4: **Trace plots for π.** The first row shows the trace for Gibbs sampling, the second row for parallel tempering, and the third row for tempered transitions. Each color represents a different component. The dashed lines indicate the true parameter values.

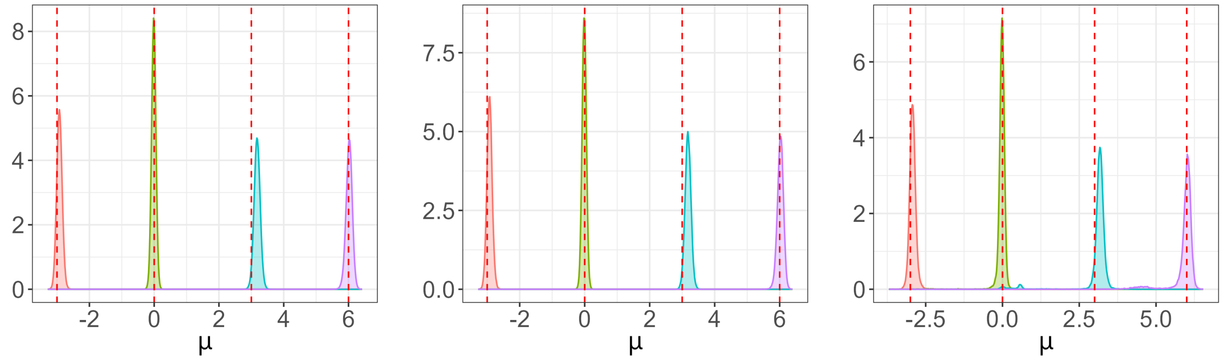

Figure 5: **Posterior distributions for μ.** The first column shows the posterior for Gibbs sampling, the second column for parallel tempering, and the third column for tempered transitions. Each color represents a different component. The dashed lines indicate the true parameter values.

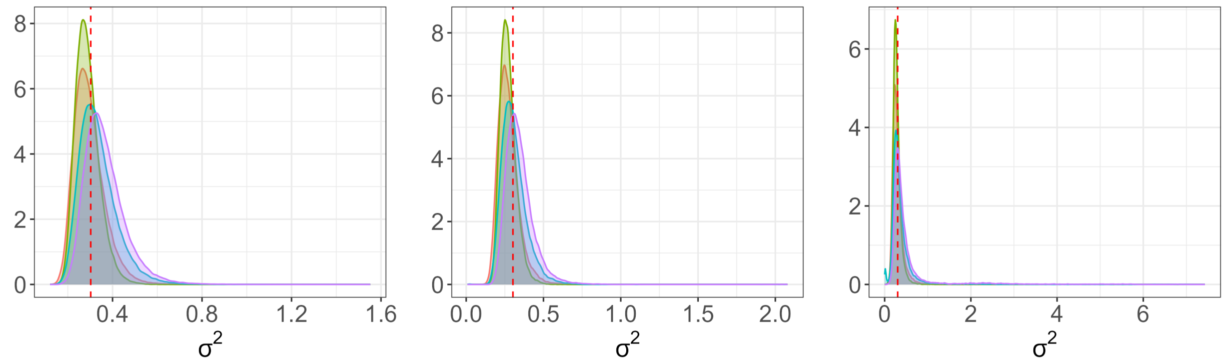

Figure 6: **Posterior distributions for $\boldsymbol{\sigma}^2$.** The first column shows the posterior for Gibbs sampling, the second column for parallel tempering, and the third column for tempered transitions. Each color represents a different component. The dashed lines indicate the true parameter values.

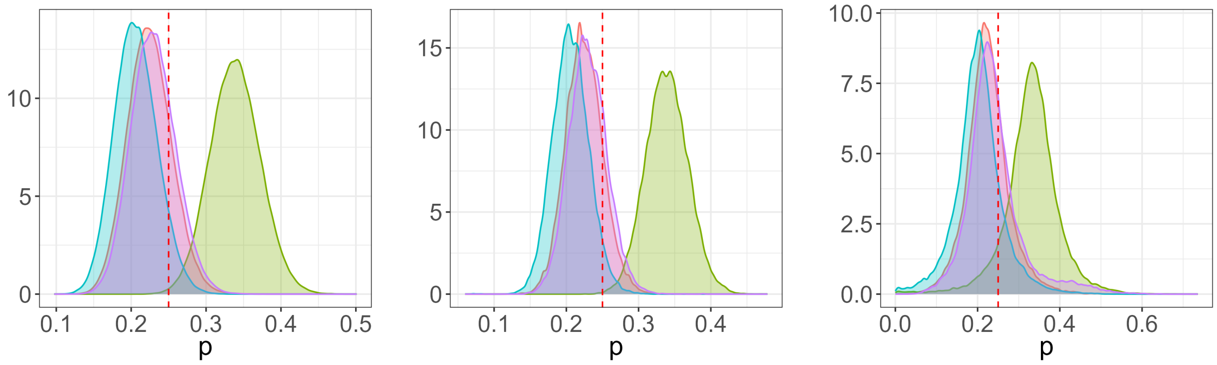

Figure 7: **Posterior distributions for $\boldsymbol{\pi}$.** The first column shows the posterior for Gibbs sampling, the second column for parallel tempering, and the third column for tempered transitions. Each color represents a different component. The dashed lines indicate the true parameter values.

References

David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7:3910–3916, 2005. doi: 10.1039/B509983H. URL `http://dx.doi.org/10.1039/B509983H`.

Charles J Geyer and Elizabeth A Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431): 909–920, 1995.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1):50 – 67, 2005. doi: 10.1214/088342305000000016. URL `https://doi.org/10.1214/088342305000000016`.

Radford M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6(4):353–366, 1996.

Matthew Sperrin, Thomas Jaki, and Ernst Wit. Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing*, 20 (3):357–366, 2010. doi: 10.1007/s11222-009-9129-8. URL `https://doi.org/10.1007/s11222-009-9129-8`.

Matthew Stephens. Bayesian methods for mixtures of normal distributions. *PhD thesis, University of Oxford*, 1997.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(4):795–809, 01 2002. ISSN 1369-7412. doi: 10.1111/1467-9868.00265. URL `https://doi.org/10.1111/1467-9868.00265`.