Annealed Importance Sampling Reversible Jump MCMC algorithms

Georgios Karagiannis*
Computational Sciences & Mathematics Division
Pacific Northwest National Laboratory
902 Battelle Boulevard, PO Box 999 MSIN: K7-90
Richland, WA 99352, USA
and
Christophe Andrieu†
Department of Mathematics
University of Bristol
University Walk, Clifton
Bristol BS8 1TW, UK

April 24, 2013

Abstract

We develop a methodology to efficiently implement the reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithms of Green (1995), applicable for example to model selection inference in a Bayesian framework, which builds on ideas of Neal (2001, 2004). We call such algorithms annealed importance sampling reversible jump (aisRJ). The proposed procedures can be thought of as being exact approximations of idealized RJ algorithms which in a model selection problem would sample the model labels only, but cannot be implemented. Central to the methodology is the idea of bridging different models with fictitious intermediate models, whose role is to introduce smooth inter-model transitions, and as we shall see improve performance. Efficiency of the resulting algorithms is demonstrated on two standard model selection problems and we show that despite the additional computational effort incurred, the approach can be highly competitive computationally. Supplemental materials for the article are available online.

Keywords: Reversible jump MCMC, annealed importance sampling, pseudo-marginal MCMC, Bayesian model selection/determination, Gaussian mixture models, Poisson change point problem.

*Georgios.Karagiannis@pnnl.gov
†C.Andrieu@bristol.ac.uk

1 Introduction

It will soon be 20 years since reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithms have been proposed by Green (1995). They have significantly extended the scope of Markov chain Monte Carlo simulation methods, offering the promise to be able to routinely tackle transdimensional sampling problems, as encountered in Bayesian model selection problems for example, in a principled and flexible fashion. Their practical efficient implementation, however, still remains a challenge. A particular difficulty encountered in practice is in the choice of the dimension matching variables (both their nature and their distribution) and the reversible transformations which allow one to define the one-to-one mappings underpinning the design of these algorithms. Indeed, even seemingly sensible choices can lead to algorithms with very poor performance. The focus of this paper is the development and performance evaluation of a method, annealed importance sampling RJ-MCMC (aisRJ), which addresses this problem by mitigating the sensitivity of RJ-MCMC algorithms to the aforementioned poor design. As we shall see the algorithm can be understood as being an "exact approximation" (Andrieu et al., 2010; Andrieu and Roberts, 2009) of an idealized MCMC algorithm that would sample from the model probabilities directly in a model selection set-up. Such an idealized algorithm may have good theoretical convergence properties, but typically cannot be implemented, and our algorithms can approximate the performance of such idealized algorithms to an arbitrary degree while not introducing any bias for any degree of approximation. Our approach combines the dimension matching ideas of RJ-MCMC with annealed importance sampling (Jarzynski, 1997b,a; Neal, 2001) and its Markov chain Monte Carlo implementation (Neal, 2004, 1996). We illustrate the performance of the algorithm with numerical simulations which indicate that, although the approach may at first appear computationally involved, it is in fact competitive.

The paper is organized as follows. In Section 2 we first introduce the notation required for the dimension matching underpinning the design of RJ-MCMC and our extension, and briefly discuss some difficulties involved in the efficient implementation of RJ-MCMC. In Section 3 we then move on to the idea of introducing intermediate (artificial) models between existing models and the subsequent idea of using importance sampling with non-homogeneous MCMC algorithms in a transdimensional set-up. This effectively extends AIS to the transdimensional set-up via ideas borrowed from the design of RJ-MCMC. This allows us in Section 4 to develop a practical MCMC implementation of AIS as suggested in (Neal, 2004), leading to AIS-RJMCMC. We conclude with a numerical investigation in Section 5.

2 Difficulties with RJ-MCMC algorithms

We consider a situation where one is interested in sampling from a probability distribution $\pi(n, \mathrm{d}\theta_n)$ defined on a union of spaces $\cup_{k\in\mathcal{K}}\{k\} \times \Theta_k$ where the spaces Θ_k may be of a different nature (e.g. dimension) and $\mathcal{K} \subset \mathbb{N}$. We will throughout the paper refer to (k, θ_k) as model k by reference to the Bayesian model selection problem, but it should be clear that the work described below is not restricted to this framework. In order to simplify presentation we will assume hereafter that for all $k \in \mathcal{K}$, $\Theta_k \subset \mathbb{R}^{\ell_k}$ for some $\ell_k \in \mathbb{N}$ and that for any $k \in \mathcal{K}$ the conditional distribution $\pi(\mathrm{d}\theta_k|k)$ has a density with respect to the Lebesgue measure on $\mathbb{R}^{\ell_k}$. Assume for now that our primary interest is in computing quantities dependent on the model probability distribution $\pi(n)$ only. If this probability distribution was available up to an unknown normalizing constant, one could use the standard Metropolis-Hastings (MH) algorithm to sample from it. Given a family of proposal distributions $\{q(k, \cdot), k \in \mathcal{K}\}$ the MH transition probability can be algorithmically described as follows

Algorithm 1 MH transition probability between model

STEP 1 Given k, sample $k'|k \sim q(k, \cdot)$.

STEP 2 Accept the transition to k' with probability $\min\{1, r_{k\to k'}\}$ where

$$r_{k\to k'} := \frac{\pi(k')q(k', k)}{\pi(k)q(k, k')} \ .$$

This algorithm is guaranteed to converge under mild assumptions, but cannot however be implemented in most scenarios of interest for which $\pi(n)$ is intractable. We will therefore refer to this algorithm as the "idealized algorithm" throughout the paper. Extending the standard Metropolis-Hastings algorithm to sample from $\pi(n, \mathrm{d}\theta_n)$ poses measure theoretic problems pertaining to the fact that $\pi(\mathrm{d}\theta_k|k)$ and $\pi(\mathrm{d}\theta_{k'}|k')$ may be defined on spaces of different nature and/or dimension. The reversible jump methodology addresses these issues with the idea of dimension matching and the introduction of reversible transformations, which allow one to embed the transdimensional problem of sampling from $\pi(n, \mathrm{d}\theta_n)$ into a family of non-transdimensional ones, while leaving the features of interest of the initial problem unchanged.

The first important ingredient of the RJMCMC methodology consists for pairs $(k, k') \in \mathcal{K}^2$ of extending the spaces $\varTheta_k$ and $\varTheta_{k'}$ to the augmented spaces $\varTheta_k \times \mathcal{U}_{k\to k'}$ and $\varTheta_{k'} \times \mathcal{U}_{k'\to k}$ with associated variables $u_{k\to k'} \in \mathcal{U}_{k\to k'} \subset \mathbb{R}^{\ell_{k\to k'}}$ and $u_{k'\to k} \in \mathcal{U}_{k'\to k} \subset \mathbb{R}^{\ell_{k'\to k}}$ for $\ell_k, \ell_{k'} \in \mathbb{N}$ such that $\ell_{k'} + \ell_{k'\to k} = \ell_k + \ell_{k\to k'}$. This embedding of the initial spaces allows one to define one-to-one mappings between the completed spaces. We denote these one-to-one mappings $G_{k\to k'} : \varTheta_k \times \mathcal{U}_{k\to k'} \to \varTheta_{k'} \times \mathcal{U}_{k'\to k}$, define

$$(\theta_{k'}(\theta_k, u_{k\to k'}), u_{k'\to k}(\theta_k, u_{k\to k'})) := G_{k\to k'}\left(\theta_k, u_{k\to k'}\right) ,$$

and we will denote $G_{k'\to k} = G^{-1}_{k\to k'}$ its inverse. The initial spaces being adequately completed one extends the initial probability distributions $\pi(k, \mathrm{d}\theta_k)$ and $\pi(k', \mathrm{d}\theta'_{k'})$ to $\pi(k, \mathrm{d}\theta_k)\varphi_{k\to k'}(\mathrm{d}u_{k\to k'})$ and $\pi(k', \mathrm{d}\theta'_{k'})\varphi_{k'\to k}(\mathrm{d}u_{k'\to k})$ for some probability distributions $\varphi_{k\to k'}(\mathrm{d}u_{k\to k'})$ and $\varphi_{k'\to k}(\mathrm{d}u_{k'\to k})$ with densities $\varphi_{k\to k'}(u_{k\to k'})$ and $\varphi_{k'\to k}(u_{k'\to k})$ with respect to the Lebesgue measure on each space. We will always assume that $\varphi_{k\to k'}(u_{k\to k'})$ and $\varphi_{k'\to k}(u_{k'\to k})$ can be evaluated and in particular that their normalizing constants are known. The two augmented spaces along with their associated variables are now explicitly associated through the functions $G_{k\to k'}$ and $G_{k'\to k}$. We will denote

$$J_{k\to k'}(\theta_k, u_{k\to k'}) := \left|\det\left(\frac{\partial G_{k\to k'}(\theta_k, u_{k\to k'})}{\partial(\theta_k, u_{k\to k'})}\right)\right| ,$$

the Jacobian of the transformation $G_{k\to k'}$. A standard RJ-MCMC update is described in (ALG2). Note that within dimension updates (i.e. $k' = k$) may differ slightly - we do not develop this here since this is standard material.

Algorithm 2 RJ-MCMC standard update

STEP 1 Given (k, θ_k), sample $k'|k \sim q(k, \cdot)$ and $u_{k\to k'} \sim \varphi_{k\to k'}(\cdot)$.

STEP 2 Accept the transition to $(k', \theta_{k'}(\theta_k, u_{k\to k'}))$ with probability $\min\{1, r_{k\to k'}\}$ where

$$r_{k\to k'} := \frac{\pi(k', \theta_{k'}(\theta_k, u_{k\to k'}))\varphi_{k'\to k}(u_{k'\to k}(\theta_k, u_{k\to k'}))q(k', k)}{\pi(k, \theta_k)\varphi_{k\to k'}(u_{k\to k'})q(k, k')} J_{k\to k'}(\theta_k, u_{k\to k'}) .$$

Although theoretically valid under mild assumptions, that is the algorithm is ensured to sample asymptotically from the joint distribution $\pi(n, \theta_n)$, the actual efficient implementation of such updates is known to be challenging : in particular good choices of $\varphi_{k\to k'}(u_{k\to k'})$, $\varphi_{k'\to k}(u_{k'\to k})$ and $G_{k\to k'}$ may not be straightforward. For example, a value of the parameter θ_k with large density under model k may be mapped into a parameter $\theta_{k'}$ of low density under model k' with a large probability, leading to high rejections rates even in situations where the model probabilities $\pi(k)$ and $\pi(k')$ are of comparable magnitude, thus resulting in very inefficient algorithms. Some suggestions have been made in (Brooks et al., 2003) about how some of these issues may be addressed.

Another approach has been suggested in (Al-Awadhi et al., 2004), and consists of modifying the RJ-MCMC in order to "improve" $\theta_{k'}(\theta_k, u_{k\to k'})$ before taking a decision about whether to accept the model jump or not. More precisely the jump to $\theta_{k'}(\theta_k, u_{k\to k'})$ is followed by the exploration of $\pi(\mathrm{d}\theta_{k'}|k')$ with T iterations of a fixed dimension and time-homogeneous MCMC algorithm targeting this distribution (or a tempered version of it) before taking any decision as to whether one should accept a transdimensional transition or not. This is clearly a sensible intuitive idea since we notice that the acceptance ratio of the idealized algorithm (ALG1) can, for example, be recovered from the RJ-MCMC standard update by letting $u_{k\to k'}$ coincide with $\theta_{k'}$, $G_{k\to k'}$ be the identity and $\varphi_{k\to k'}$ coincide with $\pi(\theta_{k'}|k')$. However, the practical implementation of the algorithm imposes that the acceptance probability of the transdimensional jumps does not improve as T increases, which is clearly not a satisfactory feature.

The approach we develop in this paper shares this idea of improving the sample quality before deciding on a transdimensional transition, but differs in many respects, and most significantly in the fact that our procedure is such that it converges to the idealized algorithm (ALG1) above as $T \to \infty$. This turns out to be an advantage since the idealized algorithm may have attractive convergence properties. For example, it can be shown that the acceptance probability of the marginal algorithm at stationarity is always larger than that of any RJ-MCMC algorithm, i.e. more precisely for any $k, k' \in \mathcal{K}$,

$$\int \pi(\mathrm{d}\theta_k|k)\varphi_{k\to k'}(\mathrm{d}u_{k\to k'})q(k,k')\min\{1, r_{k\to k'}\} \leq q(k,k')\min\left\{1, \frac{\pi(k')q(k',k)}{\pi(k)q(k,k')}\right\} ,$$

suggesting a faster visit of all the models. The result is straightforward to prove by application of Jensen's inequality. Related results in (Andrieu and Vihola, 2012) show in fact that the idealized algorithm is always superior to an exact approximation in terms of asymptotic variance. Note that as pointed out earlier, it is in principle possible to reach this upper bound in the RJ-MCMC framework, provided that sampling from $\pi(\theta_{k'}|k')$ and $\pi(\theta_k|k)$ is possible.

Crucially, our procedure relies on the idea of introducing a sequence of smoothly evolving intermediate artificial probabilistic models on the augmented space $\mathit{\Theta}_{k'} \times \mathcal{U}_{k' \to k}$ which thanks to the RJ-MCMC embedding will ensure a smooth transition between $\pi(\mathrm{d}\theta_k|k)$ and $\pi(\mathrm{d}\theta_{k'}|k')$ as opposed to a standard RJ-MCMC transition. The smoothness of this transition will manifest itself by the fact that as T increases the algorithm converges to the idealized algorithm (ALG1). This approach is related to annealed importance sampling which we now describe in a transdimensional scenario before moving to the description of aisRJ algorithms.

3 Annealed Importance sampling in a transdimensional setup

In this section we describe how the ideas of Jarzynski (1997b,a), later rediscovered in (Neal, 2001) who coined the term Annealed Importance Sampling (AIS), can be simply extended to the transdimensional scenario using the completion variables and mappings introduced by Green (1995) in the context of RJ-MCMC algorithms (see Section 2). This leads us to the introduction of the notion of intermediate models.

3.1 Importance sampling with Markov chains

For two models k and k' we are going to introduce two families of densities defined on the completed space $\Theta_{k'} \times \mathcal{U}_{k' \to k}$, the family of *forward annealing densities* $\{\rho_t(\theta_{k'}, u_{k' \to k}; k \to k'),\ t = 0, ..., T\}$ and the family of *backward annealing densities* $\{\rho_t(\theta_{k'}, u_{k' \to k}; k' \to k),\ t = 0, ..., T\}$ where the parameter t will be referred as *time* and $T \in \mathbb{N}$. As we shall see the role of these densities is going to be to interpolate the densities $\pi(\theta_k|k)\varphi_{k \to k'}(u_{k \to k'})$ and $\pi(\theta'_{k'}|k')\varphi_{k' \to k}(u_{k' \to k})$ in a smooth manner and fight poor choices of completion variables and mappings in standard reversible jumps. We will impose the following endpoint constraints (with $J_{k' \to k} = J_{k' \to k}(\theta_{k'}, u_{k' \to k})$)

$$\begin{aligned}
\rho_0(\theta_{k'}, u_{k' \to k}; k \to k') &\propto \pi(k, \theta_k(\theta_{k'}, u_{k' \to k}))\varphi_{k \to k'}(u_{k \to k'}(\theta_{k'}, u_{k' \to k}))J_{k' \to k}\ ; \\
\rho_T(\theta_{k'}, u_{k' \to k}; k \to k') &\propto \pi(k', \theta_{k'})\varphi_{k' \to k}(u_{k' \to k})\ ,
\end{aligned}$$

and similarly

$$\begin{aligned}\rho_0(\theta_{k'},u_{k'\to k};k'\to k)\propto&\pi(k',\theta_{k'})\varphi_{k'\to k}(u_{k'\to k})\ ;\\ \rho_T(\theta_{k'},u_{k'\to k};k'\to k)\propto&\pi(k,\theta_k(\theta_{k'},u_{k'\to k}))\varphi_{k\to k'}(u_{k\to k'}(\theta_{k'},u_{k'\to k}))J_{k'\to k}\ ,\end{aligned}$$

where $J_{k'\to k}(\theta_{k'},u_{k'\to k})$ is the Jacobian of the transformation $G_{k'\to k}$ expressed in terms of $(\theta_{k'},u_{k'\to k})$. The forward family $\{\rho_t(\cdot;k\to k'),\ t=1,...,T-1\}$ can be thought of as a path in an adequate space of probability distributions between $\pi(\cdot|k)\varphi_{k\to k'}(\cdot)$ to $\pi(\cdot|k')\varphi_{k'\to k}(\cdot)$, while the backward distribution family $\{\rho_t(\cdot;k'\to k),\ t=1,...,T-1\}$ provides a reverse path which does not have to coincide with the forward path. It is worth pointing out that as suggested by the above, the forward and backward annealing densities can be equivalently defined either in terms of $(\theta_k,u_{k\to k'})$ or $(\theta_{k'},u_{k'\to k})$ thanks to the one-to-one nature of $G_{k\to k'}$ and that we have simply chosen the latter. Note that we leave intermediate distributions unspecified for $t=1,\dots,T-1$ for now, and that two specific constructions will be discussed later in Subsection 3.3.

In addition to the densities above we will need a family of *forward annealing transition probabilities*

$$\{K_t((\theta_{k'},u_{k'\to k}),(\mathrm{d}\theta'_{k'},\mathrm{d}u'_{k'\to k});k\to k'),\ t=1,...,T-1\}\ ,$$

and a family of *backward annealing transition probabilities*

$$\{L_t((\theta_{k'},u_{k'\to k}),(\mathrm{d}\theta'_{k'},\mathrm{d}u'_{k'\to k});k'\to k),\ t=1,...,T-1\}\ ,$$

again defined on $\Theta_{k'}\times\mathcal{U}_{k'\to k}$, which are such that they leave invariant the family of forward annealing distributions $\{\rho_t(\mathrm{d}\theta_{k'},\mathrm{d}u_{k'\to k};k\to k'),\ t=0,...,T\}$ and the family of backward annealing distributions $\{\rho_t(\mathrm{d}\theta_{k'},\mathrm{d}u_{k'\to k};k'\to k),\ t=0,...,T\}$, respectively.

In order to alleviate notation, we now set $\tilde{\theta}^{(t)}_{k'}:=(\theta^{(t)}_{k'},u^{(t)}_{k'\to k})$, for all $t=0,...,T$. We now consider the following finite horizon non-homogeneous Markov chain with path $\tilde{\theta}^{(0:T-1)}_{k'}:=(\tilde{\theta}^{(0)}_{k'},...,\tilde{\theta}^{(T-1)}_{k'})$ on the augmented space $(\Theta_{k'}\times\mathcal{U}_{k'\to k})^T$, with the following joint probability distribution,

$$\mu_{k\to k'}(\mathrm{d}\tilde{\theta}^{(0:T-1)}_{k'}):=\rho_0(\mathrm{d}\tilde{\theta}^{(0)}_{k'};k\to k')\prod_{t=1}^{T-1}K_t(\tilde{\theta}^{(t-1)}_{k'},\mathrm{d}\tilde{\theta}^{(t)}_{k'};k\to k')\ , \tag{3.1}$$

that is $\tilde{\theta}^{(0)}_{k'}$ is sampled from $\rho_0(\mathrm{d}\tilde{\theta}^{(0)}_{k'};k\to k')$ and the $\tilde{\theta}^{(t)}_{k'}$'s are generated sequentially from $K_t(\tilde{\theta}^{(t-1)}_{k'},\mathrm{d}\tilde{\theta}^{(t)}_{k'};k\to k')$ for $t=1,...,T-1$. Similarly, we introduce the *backward annealing process* on the augmented space $(\mathit{\Theta}_{k'} \times \mathcal{U}_{k'\to k})^T$ with path $\tilde{\theta}_{k'}^{(T-1:0)} = \left(\tilde{\theta}_{k'}^{(T-1)}, ..., \tilde{\theta}_{k'}^{(0)}\right)$ and joint probability distribution

$$\mu_{k'\to k}(\mathrm{d}\tilde{\theta}_{k'}^{(T-1:0)}) = \rho_0(\mathrm{d}\tilde{\theta}_{k'}^{(T-1)}; k' \to k) \prod_{t=1}^{T-1} L_t(\tilde{\theta}_{k'}^{(T-t)}, \mathrm{d}\tilde{\theta}_{k'}^{(T-t-1)}; k' \to k) \ . \tag{3.2}$$

We now introduce the distribution $\nu_{k\to k'}(k, \mathrm{d}\tilde{\theta}_{k'}^{(0:T-1)})$ defined on $\mathcal{K} \times (\mathit{\Theta}_{k'} \times \mathcal{U}_{k'\to k})^T$

$$\nu_{k\to k'}(k, \mathrm{d}\tilde{\theta}_{k'}^{(0:T-1)}) := \pi(k)\mu_{k\to k'}(\mathrm{d}\tilde{\theta}_{k'}^{(0:T-1)}) \ ,$$

where $\pi(k)$ is the marginal probability distribution of model k and similarly the distribution $\nu_{k'\to k}(k', \mathrm{d}\tilde{\theta}_{k'}^{(T-1:0)})$ on $\mathcal{K} \times (\mathit{\Theta}_{k'} \times \mathcal{U}_{k'\to k})^T$ defined as

$$\nu_{k'\to k}(k', \mathrm{d}\tilde{\theta}_{k'}^{(T-1:0)}) := \pi(k')\mu_{k'\to k}(\mathrm{d}\tilde{\theta}_{k'}^{(T-1:0)}) \ ,$$

with $\pi(k')$ the marginal probability of model k'. We are now ready to introduce an importance sampling estimator of $\pi(k')/\pi(k)$ (we assume $\pi(k) > 0$) which relies on importance sampling with Markov chains. Provided that $\mu_{k'\to k}$ is absolutely continuous with respect to $\mu_{k\to k'}$ the Radon-Nikodym theorem applied to the measures $\nu_{k\to k'}(k, \cdot)$ and $\nu_{k'\to k}(k', \cdot)$ together with a straightforward algebraic manipulation lead to the following identity

$$\frac{\pi(k')}{\pi(k)} = \int_{(\mathit{\Theta}_{k'} \times \mathcal{U}_{k'\to k})^T} \frac{\mathrm{d}\nu_{k'\to k}\,(k', \cdot)}{\mathrm{d}\nu_{k\to k'}\,(k, \cdot)}(\tilde{\theta}_{k'}^{(0:T-1)})\mu_{k\to k'}(\mathrm{d}\tilde{\theta}_{k'}^{(0:T-1)}) \ , \tag{3.3}$$

which suggests the following unbiased estimator of $\pi(k')/\pi(k)$,

$$r_{k\to k'}^{(0:T-1)} := \frac{\mathrm{d}\nu_{k'\to k}\,(k', \cdot)}{\mathrm{d}\nu_{k\to k'}\,(k, \cdot)}(\tilde{\theta}_{k'}^{(0:T-1)}) \ ,$$

for $\tilde{\theta}_{k'}^{(0:T-1)}$ sampled from the non-homogeneous Markov chain $\mu_{k\to k'}(\mathrm{d}\tilde{\theta}_{k'}^{(0:T-1)})$. To fix ideas, in the particular scenarios where the forward and backward transition probabilities have a density with respect to the Lebesgue measure and common support $\Theta_{k'} \times \mathcal{U}_{k'\to k}$, this estimator can be rewritten in terms of all the densities involved, leading to the simple expression

$$r_{k\to k'}^{(0:T-1)} = \frac{\rho_0\left(\tilde{\theta}_{k'}^{(T-1)}; k' \to k\right)}{\rho_0\left(\tilde{\theta}_{k'}^{(0)}; k \to k'\right)} \frac{\prod_{t=1}^{T-1} L_t\left(\tilde{\theta}_{k'}^{(T-t)}, \tilde{\theta}_{k'}^{(T-t-1)}; k' \to k\right)}{\prod_{t=1}^{T-1} K_t\left(\tilde{\theta}_{k'}^{(t-1)}, \tilde{\theta}_{k'}^{(t)}; k \to k'\right)} \ . \tag{3.4}$$

We discuss later in Subsection 3.2, a set of conditions which lead to the existence of this quantity and a simple expression.

This particular scenario does not cover the practically interesting situation where the transition probabilities are Metropolis-Hastings updates, which explains the earlier abstract presentation. Naturally such estimators can be averaged for iid realizations of the process in order to define consistent estimators of $\pi(k')/\pi(k)$. An interesting fact, however, is that under realistic assumptions on the transition probabilities involved it is possible to show that this estimator is also consistent as $T \to \infty$, that is when the number of interpolating densities increases. This will turn out to have important implications on the properties of the aisRJ algorithm we later describe in the paper. The pseudo-code of the transdimensional AIS algorithm is presented in (ALG3).

Algorithm 3 Transdimensional AIS algorithm.

```
STEP 1 Initialization.
    Draw $\theta_k^{(0)}$ from the distribution $\pi\left(\mathrm{d}\theta_k^{(0)}|k\right)$.
STEP 2 AIS.
    Dimension matching
        Draw $u_{k\to k'}^{(0)}$ from the distribution $\varphi_{k\to k'}\left(\mathrm{d}u_{k\to k'}^{(0)}\right)$ and set $\left(\theta_{k'}^{(0)}, u_{k'\to k}^{(0)}\right) = G_{k\to k'}\left(\theta_k^{(0)}, u_{k\to k'}^{(0)}\right)$.
    Annealing procedure
        Generate $\left(\theta_{k'}^{(t)}, u_{k'\to k}^{(t)}\right) \sim K_t\left(\left(\theta_{k'}^{(t-1)}, u_{k'\to k}^{(t-1)}\right), \left(\mathrm{d}\theta_{k'}^{(t)}, \mathrm{d}u_{k'\to k}^{(t)}\right); k \to k'\right)$, for $t = 1, ..., T-1$.
    Annealing importance weight
        Compute the annealing importance weight $r_{k'\to k}^{(0:T-1)}$.
```

Although attractive, this approach assumes that it is possible to sample exactly from $\pi\left(\mathrm{d}\theta_k|k\right)\varphi_{k\to k'}\left(\mathrm{d}u_{k\to k'}\right)$, which will however not be possible in most scenarios of interest. A simple approach to address this problem could consist of running an MCMC with $\pi\left(\mathrm{d}\theta_k|k\right)$ as invariant distribution for a large number of iterations and use the last generated sample to initialize the AIS procedure above. Another elegant approach suggested in (Neal, 2004) in a different context consists of embedding the AIS procedure within another MCMC algorithm, effectively targeting the joint distribution $\nu_{k\to k'}(k, \mathrm{d}\tilde{\theta}_{k'}^{(0:T-1)})$: this is the approach we follow in Section 4. Before describing the aisRJ algorithm we first discuss a set of assumptions on the forward and backward annealing transitions as well as the interpolating densities which ensure the existence of the Radon-Nikodym derivative above and yields a simple expression for this quantity, valid in the situation where the forward and backward annealing transitions are Metropolis-Hastings kernels. We also discuss two systematic ways of constructing the forward and backward annealing distributions.

3.2 The AIS setup with symmetric & reversible MCMC transition probabilities

The two families of annealing distributions and the two families of annealing transition distributions have been so far discussed in abstract terms, and it may not be clear that such quantities can be defined, in particular in such a way that they can be used to perform importance sampling with Markov chains, leading to a practical AIS procedure. In what follows, we describe specific conditions on the annealing distributions and transition probabilities which lead to the existence of the desired Radon-Nikodym derivative which turns out to have a convenient simple expression. The conditions required are as follows:

Symmetry condition:

For all $t = 1, ..., T-1$ the pairs of transitions $K_t\left(\cdot, \cdot; k \to k'\right)$ and $L_{T-t}\left(\cdot, \cdot; k' \to k\right)$ satisfy the symmetry condition

$$K_t\left(\tilde{\theta}_{k'}, \mathrm{d}\tilde{\theta}'_{k'}; k \to k'\right) = L_{T-t}\left(\tilde{\theta}_{k'}, \mathrm{d}\tilde{\theta}'_{k'}; k' \to k\right) \ , \tag{3.5}$$

and for all $t = 0, ..., T$,

$$\rho_t\left(\tilde{\theta}_{k'}; k \to k'\right) = \rho_{T-t}\left(\tilde{\theta}_{k'}; k' \to k\right) . \tag{3.6}$$

Reversibility condition:

For all $t = 1, ..., T-1$,

$$\rho_t\left(\mathrm{d}\tilde{\theta}_{k'}; k \to k'\right) K_t\left(\tilde{\theta}_{k'}, \mathrm{d}\tilde{\theta}'_{k'}; k \to k'\right) = \rho_t\left(\mathrm{d}\tilde{\theta}'_{k'}; k \to k'\right) K_t\left(\tilde{\theta}'_{k'}, \mathrm{d}\tilde{\theta}_{k'}; k \to k'\right) \ . \tag{3.7}$$

Support condition:

We assume that for all $t = 0, ..., T-1$,

$$\rho_{t+1}(\tilde{\theta}_{k'}; k \to k') > 0 \Rightarrow \rho_t(\tilde{\theta}_{k'}; k \to k') > 0 \ . \tag{3.8}$$

If the annealing transition distributions $K_t(\cdot, \cdot; k \to k')$ and $L_{T-t}(\cdot, \cdot; k' \to k)$ are Metropolis-Hastings transition probabilities sharing the same proposal distributions, then condition (3.5) is automatically satisfied. Under these assumptions one can show the following crucial result.

Theorem 1. *If the symmetry condition (3.5), the reversibility conditions (3.7) and the support condition (3.8) are satisfied, then the resulting Radon-Nikodym derivative (3.3) exists. Moreover,*

$$\mu_{k' \to k}(\mathrm{d}\tilde{\theta}_{k'}^{(T-1:0)}) = \prod_{t=0}^{T-1} \frac{\rho_{t+1}\left(\theta_{k'}^{(t)}, u_{k' \to k}^{(t)}; k \to k'\right)}{\rho_t\left(\theta_{k'}^{(t)}, u_{k' \to k}^{(t)}; k \to k'\right)} \ \mu_{k \to k'}(\mathrm{d}\tilde{\theta}_{k'}^{(0:T-1)}) \ . \tag{3.9}$$

The proof is given in the appendix (available online).

Remark 2. We stress on the fact that the Conditions (3.5, 3.6 and 3.7) are not necessary for AIS or the resulting aisRJ to be valid algorithms, but that their raison d'être is that they automatically lead to the existence of the Radon-Nikodym derivative and a simple expression which can be evaluated in practice. We note also that the expression for the Radon-Nikodym derivative obtained can in fact be used directly to estimate normalizing constants without the recourse to the notion of importance sampling with Markov chains, and in particular the introduction of $\{L_t\}$. But this would not be sufficient to justify aisRJ in full generality.

3.3 Choice of the intermediate distributions.

We have until now left the annealing distributions families unspecified although they play a fundamental role in bridging the parameters of model k to those of model k'. We now describe two simple ways of systematically defining such distributions; more general can be found for example in (Gelman and Meng, 1998). Our aim here is to review two simple choices we have found useful in practice and briefly discuss some of their possible shortcomings.

3.3.1 Geometric annealing distributions

A first possibility consists of using geometric averages of the densities involved. More precisely with $\{\gamma_{t,T}, t = 0, \ldots, T\} \in [0,1]^T$ (with $\gamma_{0,T} = 0$ and $\gamma_{T,T} = 1$) one can define the geometric annealing densities $\rho_t(\theta_{k'}, u_{k'\to k}; k \to k')$ and $\rho_t(\theta_{k'}, u_{k'\to k}; k' \to k)$ as follows (with $J_{k'\to k} = J_{k'\to k}(\theta_{k'}, u_{k'\to k})$)

$$\begin{aligned}\rho_t(\theta_{k'}, u_{k'\to k}; k \to k') \propto & \left\{\pi(k, \theta_k(\theta_{k'}, u_{k'\to k}))\varphi_{k\to k'}(u_{k\to k'}(\theta_{k'}, u_{k'\to k}))\ J_{k'\to k}\right\}^{1-\gamma_{t,T}} \\ & \times \left\{\pi(k', \theta_{k'})\varphi_{k'\to k}(u_{k'\to k})\right\}^{\gamma_{t,T}} \ ,\end{aligned} \tag{3.10}$$

and

$$\begin{aligned}\rho_t(\theta_{k'}, u_{k'\to k}; k' \to k) \propto & \left\{\pi(k, \theta_k(\theta_{k'}, u_{k'\to k}))\varphi_{k\to k'}(u_{k\to k'}(\theta_{k'}, u_{k'\to k}))\ J_{k'\to k}\right\}^{1-\gamma_{T-t,T}} \\ & \times \left\{\pi(k', \theta_{k'})\varphi_{k'\to k}(u_{k'\to k})\right\}^{\gamma_{T-t,T}} \ ,\end{aligned} \tag{3.11}$$

respectively, for $t = 0, ..., T$. It should be clear that $\varphi_{k\to k'}$ and $\varphi_{k'\to k}$ should be such that they let $u_{k\to k'}$ or $u_{k'\to k}$ evolve freely in order for the regions of high density under $\pi(k', \theta_{k'})$ to be easily reached, and we note that it is, in principle, possible to make these densities and the transformations time dependent. In this situation and under the conditions of Theorem 1 the annealing weight takes the form (with the convention $\theta_k^{(t)} := \theta_k(\theta_{k'}^{(t)}, u_{k'\to k}^{(t)})$ and $u_{k\to k'}^{(t)} = u_{k\to k'}(\theta_{k'}^{(t)}, u_{k'\to k}^{(t)})$

$$r_{k\to k'}^{(0:T-1)} = \prod_{t=0}^{T-1} \left(\frac{\pi\left(k', \theta_{k'}^{(t)}\right)}{\pi\left(k, \theta_k^{(t)}\right)} \frac{\varphi_{k'\to k}\left(u_{k'\to k}^{(t)}\right)}{\varphi_{k\to k'}\left(u_{k\to k'}^{(t)}\right)} J_{k'\to k}^{-1}(\theta_{k'}^{(t)}, u_{k'\to k}^{(t)}) \right)^{\gamma_{t+1,T}-\gamma_{t,T}} \ , \tag{3.12}$$

and one notices that computation of this quantity only requires one to be able to evaluate $\pi(n, \theta_n)$ up to a normalizing constant. This choice may not always lead to an efficient AIS estimator. Indeed the product form of these densities implies that, as is the case for standard RJ-MCMC algorithms, a poor choice of matching variables and mapping can lead to large variations between $\rho_0(\theta_{k'}, u_{k'\to k}; k \to k')$ and $\rho_1(\theta_{k'}, u_{k'\to k}; k \to k') \simeq 0$ for example, resulting in estimators with a large variance. This may require increasing T significantly or making the grid initially finer. Another approach, less sensitive to this phenomenon, consists of considering arithmetic means.

3.3.2 Arithmetic annealing distributions

It is important here to consider unormalized densities $g_{k'}(\theta_{k'}, u_{k'\to k}) \propto \pi(k', \theta_{k'})\varphi_{k'\to k}(u_{k'\to k})$ and $g_k(\theta_{k'}, u_{k'\to k}) \propto \pi(k, \theta_k(\theta_{k'}, u_{k'\to k}))\varphi_{k\to k'}(u_{k\to k'}(\theta_{k'}, u_{k'\to k}))J_{k'\to k}(\theta_{k'}, u_{k'\to k})$ that we are able to evaluate in practice. The arithmetic forward annealing distribution densities are defined on $\Theta_{k'} \times \mathcal{U}_{k'\to k}$ as

$$\rho_{t,T}(\theta_{k'}, u_{k'\to k}; k \to k') = \frac{(1-\gamma_{t,T})g_k(\theta_{k'}, u_{k'\to k}) + \gamma_{t,T}g_{k'}(\theta_{k'}, u_{k'\to k})}{(1-\gamma_{t,T})\pi(k) + \gamma_{t,T}\pi(k')} \ , \tag{3.13}$$

for $t = 0, ..., T$. Since no product is involved, the problem inherent to the geometric mean approach seems to have disappeared. One should however be cautious. Due to the fact that we are in practice forced to use unormalized densities, the true nature of (3.13) as a density on $\varTheta_{k'} \times \mathcal{U}_{k'\to k}$ is obtained by rewriting

$$\begin{aligned}\rho_{t,T}(\theta_{k'}, u_{k'\to k}; k \to k') = &\\ &\frac{(1-\gamma_{t,T})\pi(k)}{(1-\gamma_{t,T})\pi(k) + \gamma_{t,T}\pi(k')}\left(\frac{g_k(\theta_{k'}, u_{k'\to k})}{\pi(k)}\right)\\ &+ \frac{\gamma_{t,T}\pi(k')}{(1-\gamma_{t,T})\pi(k) + \gamma_{t,T}\pi(k')}\left(\frac{g_{k'}(\theta_{k'}, u_{k'\to k})}{\pi(k')}\right) \ ,\end{aligned} \tag{3.14}$$

which reveals the true nature of the parametrization introduced. In particular, one notices that $\rho_{t,T}(\theta_{k'}, u_{k'\to k}; k \to k')$ may not evolve smoothly, even when the increments of $\{\gamma_{t,T}\}$ are apparently small, in the particular case where $\pi(k)$ is very different from $\pi(k')$ for a given pair $(k,\ k') \in \mathcal{K}^2$. This may require choosing T large in order to ensure smoothness.

4 AISRJ MCMC

We are now ready to describe the aisRJ algorithm. Throughout this section we assume that the conditions of Theorem 1 hold. The aisRJ algorithm builds on the AIS estimator introduced in the previous section and its transition probability is described in (ALG4) (again we use the convention $\theta_k^{(0)} := \theta_k(\theta_{k'}^{(0)}, u_{k'\to k}^{(0)})$ and $u_{k\to k'}^{(0)} = u_{k\to k'}(\theta_{k'}^{(0)}, u_{k'\to k}^{(0)})$. Note that T may depend on the pair (k, k'), but we omit here this dependence in order to alleviate notation. The important feature of the algorithm is that similarly to the algorithm of Al-Awadhi et al. (2004), an MCMC algorithm is used to improve the standard RJ-MCMC initial "jump" to model k', but in such a way that it leads to a particularly interesting feature: under realistic assumptions, the acceptance probability of (ALG4) converges to that of (ALG1), and can therefore be thought of as being an approximation of this idealized algorithm. We now show that this algorithm is exact, that is it leaves $\pi(n, \theta_n)$ invariant. This is shown by proving reversibility, which is a direct consequence of the crucial result established in Theorem 1.

Algorithm 4 aisRJ transition probability

Let (k, θ_k) be the current state of the Markov chain.

STEP 1 Model proposal move.

Propose model k', with probability $q(k, \cdot)$.

STEP 2 AIS sweep.

Dimension matching

Set $\theta_k^{(0)} = \theta_k$, draw $u_{k\to k'}^{(0)} \sim \varphi_{k\to k'}(\mathrm{d}u_{k\to k'})$ and compute $\left(\theta_{k'}^{(0)}, u_{k'\to k}^{(0)}\right) = G_{k\to k'}\left(\theta_k^{(0)}, u_{k\to k'}^{(0)}\right)$.

Annealing procedure

Generate a path $\left(\theta_{k'}^{(1)}, u_{k'\to k}^{(1)}\right), ..., \left(\theta_{k'}^{(T-1)}, u_{k'\to k}^{(T-1)}\right)$, where $\left(\theta_{k'}^{(t)}, u_{k'\to k}^{(t)}\right)$ is drawn from the Markov transition distribution $K_t\left(\left(\theta_{k'}^{(t-1)}, u_{k'\to k}^{(t-1)}\right), \left(\mathrm{d}\theta_{k'}^{(t)}\mathrm{d}u_{k'\to k}^{(t)}\right); k \to k'\right)$, for some integer $T > 1$.

Annealing importance weight

Compute the annealing importance weight $r_{k\to k'}^{(0:T-1)}$,

$$r_{k\to k'}^{(0:T-1)} = \frac{\pi\left(k', \theta_{k'}^{(T-1)}\right)}{\pi\left(k, \theta_k^{(0)}\right)} \frac{\varphi_{k'\to k}\left(u_{k'\to k}^{(T-1)}\right)}{\varphi_{k\to k'}\left(u_{k\to k'}^{(0)}\right)} J_{k'\to k}^{-1}(\theta_{k'}^{(0)}, u_{k'\to k}^{(0)}) \\ \times \prod_{t=1}^{T-1} \frac{\rho_t\left(\theta_{k'}^{(t-1)}, u_{k'\to k}^{(t-1)}; k \to k'\right)}{\rho_t\left(\theta_{k'}^{(t)}, u_{k'\to k}^{(t)}; k \to k'\right)} \, . \tag{4.1}$$

STEP 3 RJ: Accept/reject step

Accept the proposed value $\left(k', \theta_{k'}^{(T-1)}\right)$ with acceptance probability:

$$a_{k,k'}^{(0:T-1)} = \min\left\{1, \frac{q(k',k)}{q(k,k')} r_{k\to k'}^{(0:T-1)}\right\} \, . \tag{4.2}$$

Theorem 3. *Under the conditions of Theorem 1, namely (3.5, 3.6), (3.7) and (3.8), the aisRJ algorithm in (ALG4) is reversible with respect to* $\pi(n, \theta_n)$*.*

The proof is given in the appendix (available online).

The ancestry of this algorithm can be traced back, to the best of our knowledge, to (Neal, 1996) where the purpose was primarily to remove the random walk behavior of tempering methods and the computation of normalizing constants a by-product used to justify the asymptotic expected performance of the algorithm. The use of importance sampling with Markov chains for the computation of normalizing constants was later rediscovered in (Jarzynski, 1997b,a) and (Neal, 2001), but it is the algorithm in (Neal, 2004) which is the clear direct (and close) ancestor of the present procedure.

5 Examples

In order to illustrate the properties and performance of the aisRJ approach, we apply the methodology to a toy example for purely pedagogical purposes, the Poisson multiple change problem as investigated in (Green, 1995) and the classical finite Gaussian mixture model determination/selection problem as addressed in (Richardson and Green, 1997). Other examples have been considered in (Karagiannis, 2011) and similar conclusions were drawn. The main aim of our evaluation is to investigate the dependence of the performance of the algorithm on T (the number of intermediate distributions), after N iterations : this includes the standard RJ-MCMC (stdRJ) described in (ALG4) for the case $T = 1$ and the idealized algorithm (idlRJ) described in (ALG1), which corresponds to the situation $T = \infty$. In the two examples, we have considered that $\pi(n)$ is not available to implement (ALG1), and we have therefore considered the simple proxy which consists of using estimates $\hat{\pi}(n)$ of $\pi(n)$ obtained from long runs of the other methods. For brevity, and since the two examples heavily rely on the modeling and algorithmic ideas of (Green, 1995; Richardson and Green, 1997), we mainly focus on the algorithmic differences in our implementation and the analysis of the performance and refer the reader to those papers for details.

5.1 A toy example

Similarly to Andrieu and Roberts (2009), we assume that it is of interest to sample from the transdimensional distribution $\pi(n, \theta)$ with density

$$\begin{aligned}\pi(n,\theta) =& \frac{1}{4}\mathcal{N}(\theta=\theta_1;\mu_1=0,\Sigma_1=1)\mathbb{I}_{\{1\}\times\mathbb{R}}(n,\theta)\\ &+\frac{3}{4}\mathcal{N}_2\left(\theta=\theta_2;\mu_2=\begin{pmatrix}0\\0\end{pmatrix},\Sigma_2=\begin{pmatrix}1 & -0.9\\ -0.9 & 1\end{pmatrix}\right)\mathbb{I}_{\{2\}\times\mathbb{R}^2}(n,\theta),\end{aligned} \tag{5.1}$$

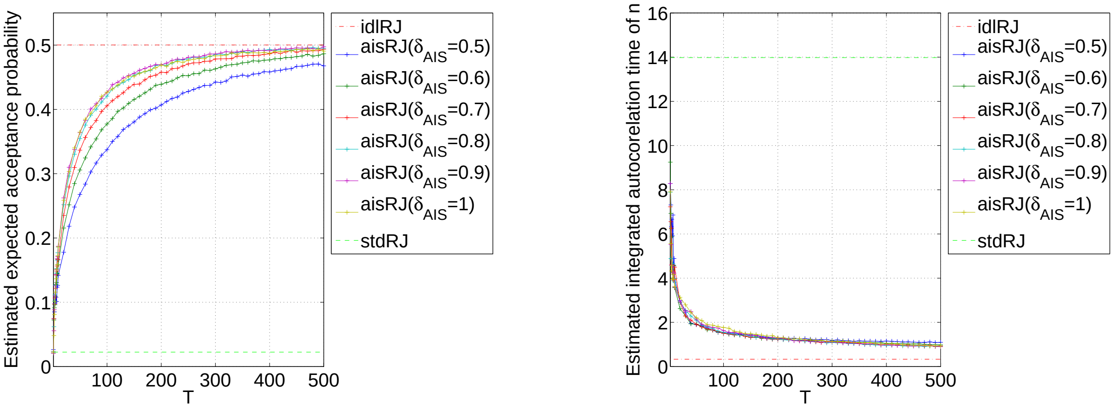
defined on the space $(\{1\}\times\mathbb{R})\cup(\{2\}\times\mathbb{R}^2)$. The marginal model posterior probability for models is $\pi(n)=\frac{1}{4}\mathbb{I}_{\{1\}}(n)+\frac{3}{4}\mathbb{I}_{\{2\}}(n)$ and the conditional posterior distributions for their within model parameters are $\pi(\theta=\theta_1|n=1)=\mathcal{N}(\theta_1;\mu_1,\Sigma_1)$, $\pi(\theta=\theta_2|n=2)=\mathcal{N}_2\left(\theta_2;\mu_2,\Sigma_2\right)$ respectively.

A simple idlRJ algorithm (ALG1) that targets $\pi(n)$ could use marginal proposal distribution $q(1,2)=q(2,1)=1$ if $\pi(n)$ was tractable. Here, we pretend that $\pi(n)$ is intractable and $\pi(n,\theta)$ is known up to a common normalizing constant. In this case, a standard approach would be to resort to a RJ-MCMC algorithm that targets the joint distribution $\pi(n,\theta)$. We will consider here a stdRJ algorithm with marginal proposal distribution $q(1,2)=q(2,1)=1$, dimension matching proposal distribution $\varphi_{1\to2}(u_{1\to2})=\mathcal{N}(u_{1\to2};3,1)$ and transformation function the identity, i.e. $\theta_2=G_{1\to2}(\theta_1,u_{1\to2})=(\theta_1,u_{1\to2})$. Although such a stdRJ converges in theory, the chosen reversible jump proposals are a poor choice because the proposed values may often lie far from the mode of $\pi(2,\theta_2)$ when a move $1\to2$ attempts.

In order to improve the stdRJ algorithm, we design an aisRJ (ALG4) based on the aforementioned reversible jump proposals. We consider a family of forward geometric annealing distributions $\rho_t(\mathrm{d}\theta_2;1\to2)$, defined on the parameter space $\mathbb{R}^2$ of model 2, with density

$$\begin{aligned}\rho_t(\theta_2=(\theta_{2,1},\theta_{2,2});1\to2)\propto&\left\{\frac{1}{4}\mathcal{N}(\theta_{2,1};\mu_1,\Sigma_1)\times\mathcal{N}(\theta_{2,2};3,1)\times1\right\}^{\frac{t-T}{T}}\\ &\times\left\{\frac{3}{4}\mathcal{N}_2\left((\theta_{2,1},\theta_{2,2});\mu_2,\Sigma_2\right)\right\}^{\frac{t}{T}},\end{aligned}$$

which in the present special case is a sequence of weighted normal distributions with means and variance interpolating $(\mu_1,3)$ and μ_2, and $\mathrm{diag}(\Sigma_1,3)$ and Σ_2. We take the forward annealing transition probabilities $\{K_t(\cdot,\mathrm{d}\cdot),\ t=1,...,T\}$ to be MALA Metropolis-Hastings updates (Roberts and Rosenthal, 1998) targeting $\{\rho_t(\mathrm{d}\cdot;1\to2),\ t=1,...,T\}$ with scale parameter δ_{AIS}. Namely, the Metropolis-Hastings updates target $\{\rho_t(\theta_2;1\to2),\ t=1,\ldots,T\}$ using proposals $\{\mathcal{N}_2(\theta_2';\theta_2+\frac{\delta_{\mathrm{AIS}}}{2}\nabla\log\rho_t(\theta_2;1\to2),\ \delta_{\mathrm{AIS}}),\ t=1,...,T\}$. The backward annealing proposal proba-



(a) Expected acceptance probability

(b) Integrated autocorrelation time

Figure 5.1: Estimated expected acceptance probabilities and integrated autocorrelated times as functions of the total annealing time T. The involved algorithms in the plots are aisRJ, idlRJ and stdRJ. The estimates are computed after $N = 10^5$ iterations of the aisRJ update. The involved scaling values δ_{AIS} for the annealing proposals of aisRJ are 0.5, 0.6, 0.7, 0.8, 0.9 and 1.

bilities $\{L_t(\cdot, \mathrm{d}\cdot),\ t = 1, ..., T\}$ are such that $L_t = K_{T-t}$ for $t = 1, ..., T$. The annealed importance weight can be found to be

$$r_{1\to 2}^{(0:T-1)} = \prod_{t=0}^{T-1} \left(\frac{3\mathcal{N}_2\left((\theta_{2,1}^{(t)}, \theta_{2,2}^{(t)})^{\mathrm{T}}; \mu_2, \Sigma_2\right)}{\mathcal{N}(\theta_{2,1}^{(t)}; \mu_1, \Sigma_1)} \times \frac{1}{\mathcal{N}(\theta_{2,2}^{(t)}; 3, 1)} \times 1 \right)^{\frac{1}{T}} ,$$

for the move $1 \to 2$ and $r_{2\to 1}^{(0:T-1)} = 1/r_{1\to 2}^{(0:T-1)}$ for the move $2 \to 1$.

In our numerical example, we ran aisRJ for $N = 10^5$ iterations and total annealing times T in the range $\{1, ..., 500\}$. A reasonable scaling parameter for the annealing proposals is $\delta_{\text{AIS}} = 0.8$ according to Roberts and Rosenthal (1998). We observe that the expected acceptance probability increases with T (Fig. 5.1a) while the integrated autocorrelation time of index n decreases when T increases (Fig. 5.1b). For large enough T, aisRJ is observed to converge to idlRJ in terms of the expected acceptance probability and integrated autocorrelation time of n. The plots suggest that the aisRJ implementation of RJ-MCMC improves significantly on the poorly designed stdRJ caused by the badly chosen RJ proposals $\varphi_{1\to 2}(\cdot)$ and values of δ_{AIS}.

5.2 Poisson multiple change point models

We consider the analysis of the *coal-mining disasters* dataset by using the model proposed in (Green, 1995). Hence, following (Green, 1995) we assume that the n data points $\{y_i,\ i = 1, ..., n\}$ (times of occurrence of disasters and we set $y_0 = 0$) arise from a non-homogeneous Poisson process model on a time interval $[0, L]$ with intensity $x(\cdot)$ modeled as a step function. The number of steps, $k + 1$, is unknown, $(k \in \mathcal{K} = \{0, ..., k_{\max}\})$ and we denote the starting points for each step as $\{s_{j,k},\ j = 1, \ldots, k\}$, with the constraint $0 = s_{0,k} < s_{1,k} < ... < s_{k+1,k} = L$. The step function takes the value $h_{j,k}$, referred to as the height, for the segment $[s_{j,k}, s_{j+1,k})$, for $j = 0, ..., k$. We denote the random parameter vector of model k as $\phi_k = (s_{j,k}, h_{j,k},\ j = 0, ..., k)$. As a result, the likelihood of the model k is

$$\log\left(\mathcal{L}_k\left(y_{1:n}|\phi_k\right)\right) = \sum_{i=1}^{n} \log\left(x_k\left(y_i;\phi_k\right)\right) - \int_0^L x_k\left(t;\phi_k\right) \mathrm{d}t\ , \tag{5.2}$$

where $x_k(t;\phi_k) = \sum_{j=0}^{k} h_{j,k}\mathbb{I}_{[s_{j,k},s_{j+1,k})}(t)$, for $t \in [0, L]$. We use the priors and hyperparameters suggested by Green (1995) which involves the introduction of hyperparameters α_k, β_k for the prior distribution of the heights within model k (a Gamma distribution), which form part of the inference. As a result in the present example $\theta_k := (\phi_k, \alpha_k, \beta_k)$ and $\Theta_k = [0, L]^k \times (0, +\infty)^{k+1} \times (0, +\infty) \times (0, +\infty)$.

We build our aisRJ around the pair of reversible jump updates proposed in (Green, 1995), which consist of a step split and a steps merge moves. The first move splits one segment into two neighbouring segments by adding one change point at location ν drawn uniformly at random in $[0, L]$ and distributing the corresponding level between the two new steps according to the formulae of Green (1995). The second move combines two neighboring segments drawn uniformly at random, segments number $u_{k+1\to k} = j^*_{k+1}$ and $j^*_{k+1} + 1$, into one segment by removing the change point at their boundaries and inverting the formulae used in the split move. As a result, one has $\tilde{\theta}_{k+1} = (\theta_{k+1}, j^*_{k+1})$, defined on extended space and $\tilde{\Theta}_{k+1} = \Theta_{k+1} \times \{0, ..., k\}$. The proposal distributions $\{q(k, \cdot);\ k \in \mathcal{K}\}$ are as in (Green, 1995). We have considered the geometric mean type intermediate distributions with $\gamma_{t,T} = t/T$ leading to

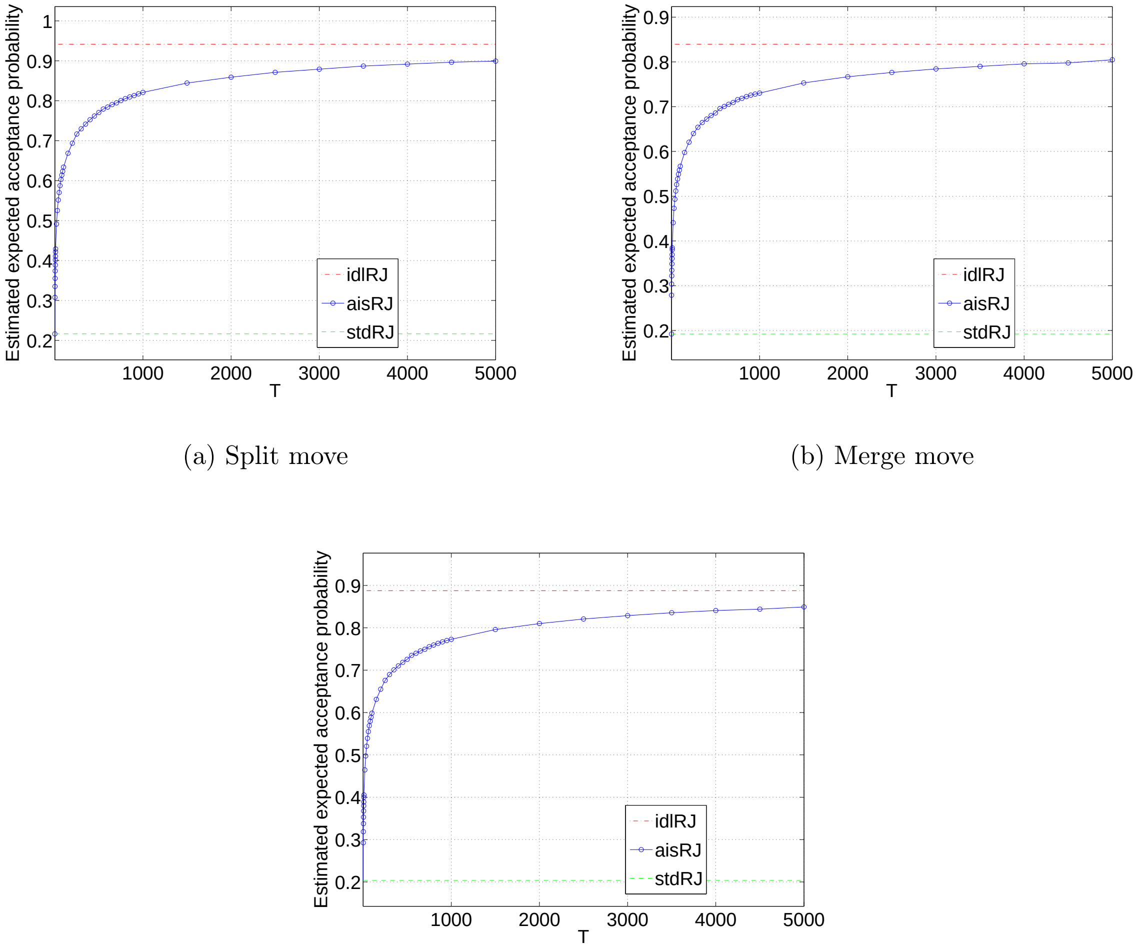
$$\rho_t\left(\tilde{\theta}_{k+1}; k \to k+1\right) \propto \left(\pi\left(k, \theta_k | y\right) \frac{1}{L} \frac{\left(h_{j^*_{k+1},k+1} + h_{j^*_{k+1}+1,k+1}\right)^2}{h_{j^*_{k+1},k}}\right)^{\frac{T-t}{T}} \times \left(\pi\left(k+1, \theta_{k+1} | y\right) \frac{1}{k+1}\right)^{\frac{t}{T}}, \tag{5.3}$$

for $t = 0, ..., T$.

It is important to ensure that the required reversibility (3.7) for the annealing transition probabilities $\{K_t,\ t = 0, \ldots, T\}$ is satisfied and we consider random permutation blockwise MCMC sweeps as outlined in (ALG5) (which target $\rho_t(\mathrm{d}\cdot; k \to k+1)$ for $t = 0, ..., T$). The same update is used for the within model updates, for which $T = 1$. We naturally made the choice $L_t(\cdot, \mathrm{d}\cdot; k+1 \to k) = K_{T-t}(\cdot, \mathrm{d}\cdot; k \to k+1)$ for $t = 0, ..., T$.

It is worth pointing out that sampling j^*_{k+1} is required in order to ensure that aisRJ converges to idlRJ as $T \to \infty$ and that although we have focus throughout the paper on approximating idlRJ, one may alternatively target partially marginalized algorithms. In Fig. 5.2 we report the expected acceptance probabilities of the two updates (conditionally or unconditionally on the type of move), and observe that as expected they increase as T increases. In particular the standard RJ-MCMC ($T = 1$) performs worse, while idlRJ ($T = \infty$) performs best.

In Fig. 5.3 we report the autocorrelation function of k for increasing values of T and observe again the benefit of increasing T (omitting for now the additional computational overhead). The corresponding integrated autocorrelation times are reported as a function of T in Fig. 5.4c. The striking and important feature which we have repeatedly observed in our experiments is the initial sharp drop in the value of the integrated autocorrelation time as T increases (for $T \leq T_0$ for some $T_0 \in \mathbb{N}$), at a rate clearly faster than the standard Monte Carlo rate of T^{-1}, which is then followed by a much slower rate of improvement of the order $O(T^{-1})$. This suggests to us that T_0 is related to the difficulty of transiting from models to models, the problem precisely addressed by the AIS strategy, and that once this difficulty is overcome we recover the standard Monte Carlo rate of convergence. In Fig. 5.4a and 5.4b, we observe the same phenomenon for the estimated integrated autocorrelation times for the functions $\mathbb{I}_{\{1\}}(k)$ and $\mathbb{I}_{\{4\}}(k)$ and similarly for $s_2\mathbb{I}_{[2,+\infty)}(k)$, $h_2\mathbb{I}_{[2,+\infty)}(k)$ and $s_3\mathbb{I}_{[3,+\infty)}(k)$ (renormalized to compute the corresponding conditional expectations) in Fig. 5.5.

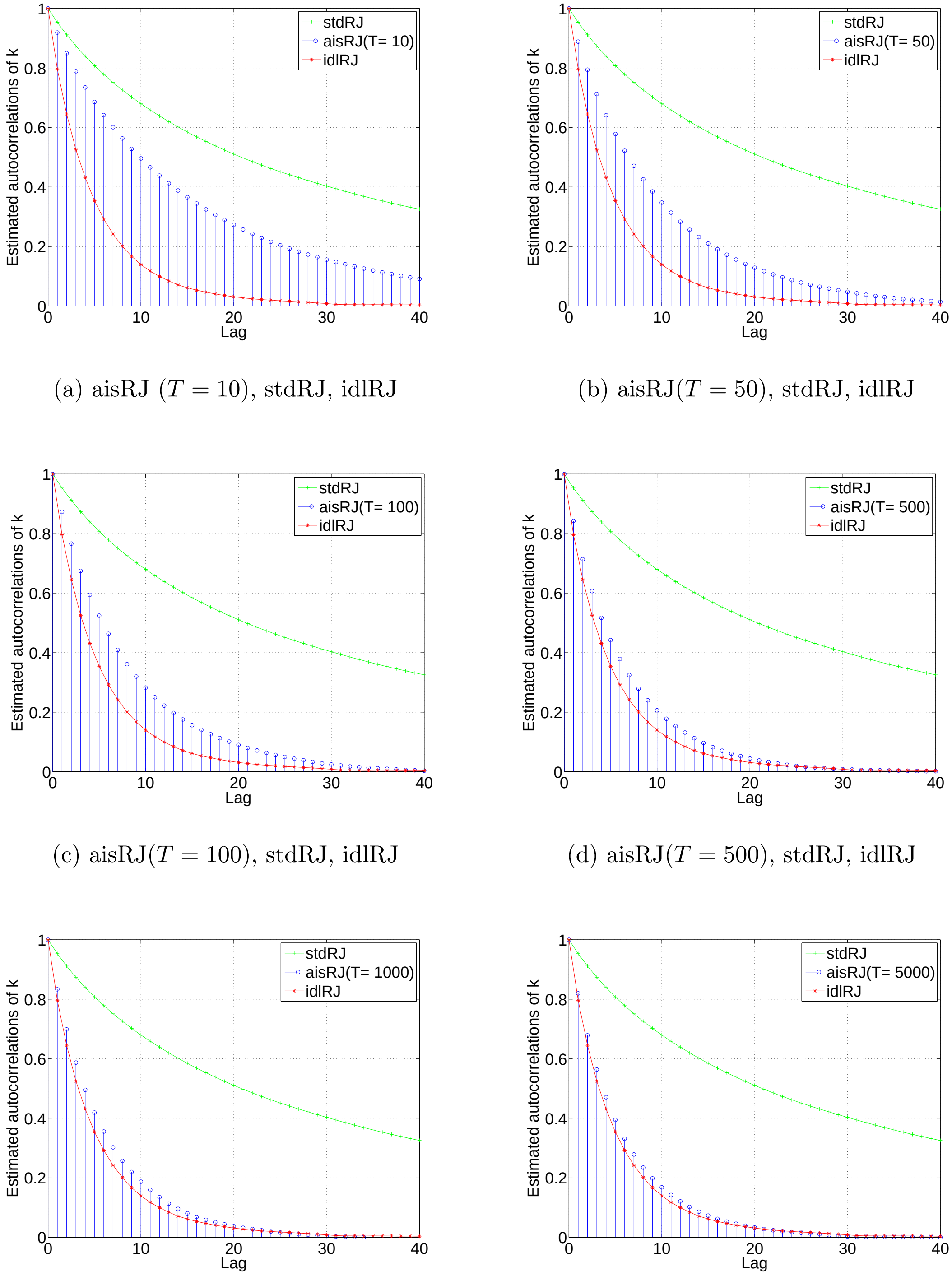


(a) Split move

(b) Merge move

(c) Split & Merge move

Figure 5.2: Estimated expected acceptance probabilities for the complete split & merge pair of moves, the split move only and the merge move only. The involved algorithms in the plot are the stdRJ (- - -), the aisRJ (–o–) and the approximated idlRJ (-·-).



(a) aisRJ ($T = 10$), stdRJ, idlRJ

(b) aisRJ($T = 50$), stdRJ, idlRJ

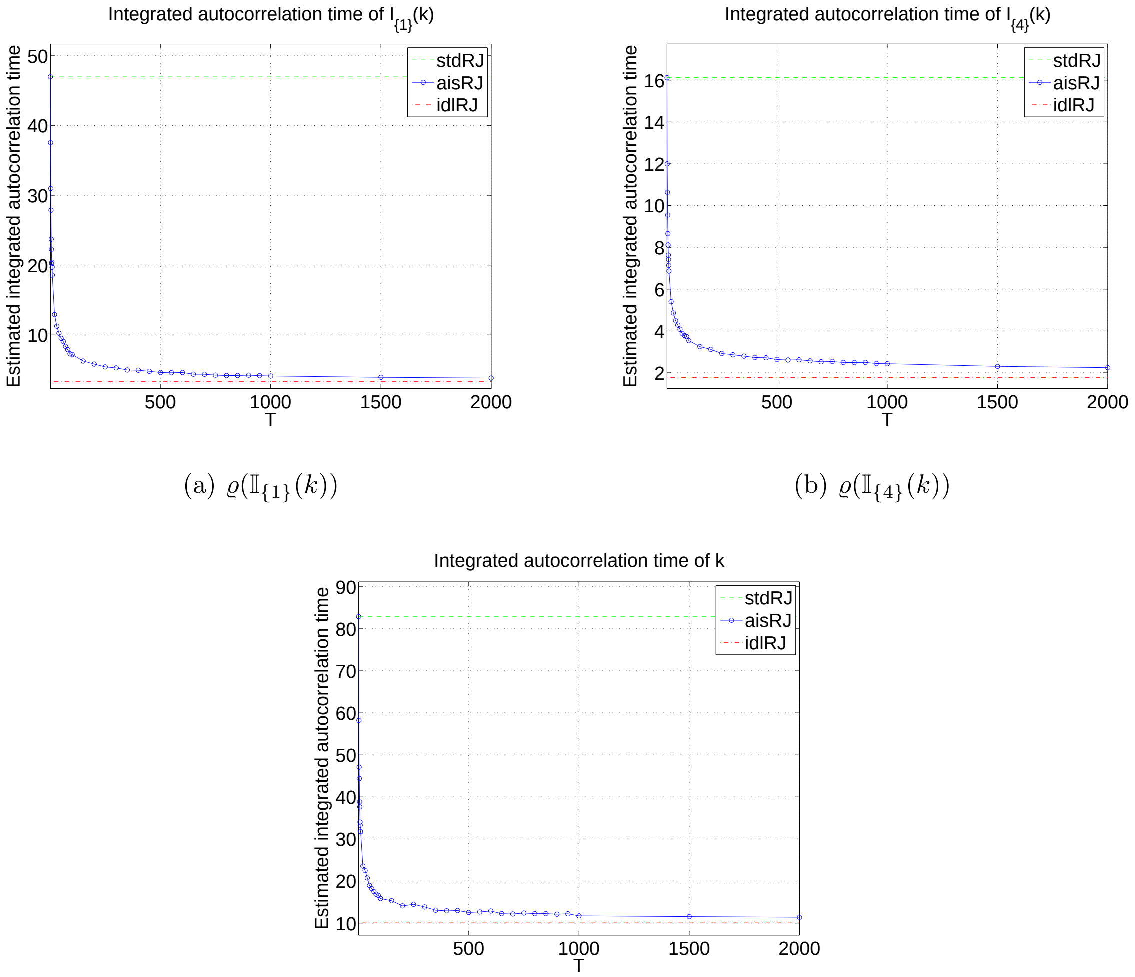
(c) aisRJ($T = 100$), stdRJ, idlRJ

(d) aisRJ($T = 500$), stdRJ, idlRJ

(e) aisRJ($T = 1000$), stdRJ, idlRJ

(f) aisRJ($T = 5000$), stdRJ, idlRJ

Figure 5.3: Autocorrelation function plot of k. The involved algorithms in the plot are the stdRJ (—), the aisRJ (−o) and the approximated idlRJ (—). The number of iterations N is equal to $5 \cdot 10^5$.

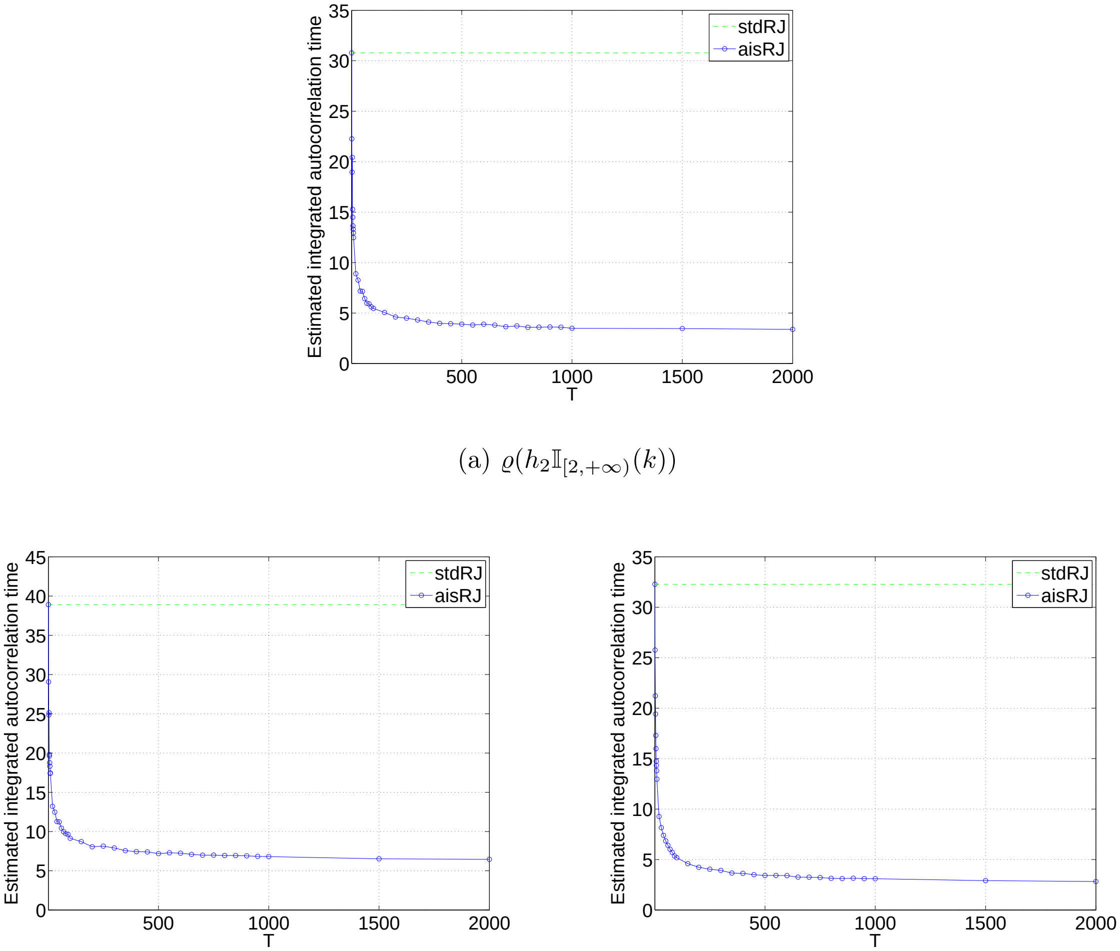


(a) $\varrho(\mathbb{I}_{\{1\}}(k))$

(b) $\varrho(\mathbb{I}_{\{4\}}(k))$

(c) $\varrho(k)$

Figure 5.4: Estimated integrated autocorrelation times $\varrho(k)$, $\varrho(\mathbb{I}_{\{1\}}(k))$ and $\varrho(\mathbb{I}_{\{4\}}(k))$. The involved algorithms in the plot are the stdRJ (- - -), the aisRJ (—o—) and the approximated idlRJ (-·-). The number of iterations N is equal to $5 \cdot 10^5$ and the total annealing time $T = \{1, ..., 2000\}$.



(a) $\varrho(h_2\mathbb{I}_{[2,+\infty)}(k))$

(b) $\varrho(s_2\mathbb{I}_{[2,+\infty)}(k))$

(c) $\varrho(s_3\mathbb{I}_{[3,+\infty)}(k))$

Figure 5.5: Estimated integrated autocorrelation times $\varrho(s_2\mathbb{I}_{[2,+\infty)}(k))$, $\varrho(s_3\mathbb{I}_{[3,+\infty)}(k))$ and $\varrho(h_2\mathbb{I}_{[2,+\infty)}(k))$. The involved algorithms in the plot are the stdRJ (- - -) and the aisRJ (−o−). The number of iterations N is equal to $5 \cdot 10^5$ and the total annealing time $T = \{1, ..., 2000\}$.

Algorithm 5 Components of the blockwise MCMC sweep.

```
• Metropolis-Hastings block which updates the $j$-th height:
    Draw $h'_{j,k+1}$ such that $\log\left(h'_{j,k+1}/h_{j,k+1}\right) \sim\mathcal{U}(-0.5, 0.5)$.
    Accept $h'_{j,k+1}$ with probability $\min\left\{1, \frac{\rho_t(h'_{j,k+1}|...;k\rightarrow k+1)}{\rho_t(h_{j,k+1}|...;k\rightarrow k+1)}\right\}$.
• Metropolis-Hastings block which updates the $j$-th change point:
    Draw $s'_{j,k+1}$ from $\mathcal{U}(s_{j-1,k+1}, s_{j+1,k+1})$
    Accept $s'_{j,k+1}$ with probability $\min\left\{1, \frac{\rho_t(s'_{j,k+1}|...;k\rightarrow k+1)}{\rho_t(s_{j,k+1}|...;k\rightarrow k+1)}\right\}$ .
• Metropolis-Hastings block which updates $\beta_{k+1}$:
    Draw $\beta'_{k+1}$from $\mathcal{G}a\left(\tilde{e}_t, \tilde{f}_t\right)$ where $\tilde{e}_t = e + (k+1+\gamma_t)\alpha_k$ and
    $\tilde{f}_t = f + \sum_{j\neq j^*,j^*+1} h_{j,k+1} + (1-\gamma_t)\,h_{j^*,k} + \gamma_t h_{j^*,k+1} + \gamma_t h_{j^*+1,k+1}$.
• Metropolis-Hastings block which updates $\alpha_{k+1}$:
    Draw $\alpha'_{k+1}$ as $\alpha'_{k+1} = \alpha_{k+1}2^{u-0.5}$, where $u \sim\mathcal{U}(0,1)$.
    Accept $\alpha'_{k+1}$ with probability $\min\left\{1, \frac{\rho_t(\alpha'_{k+1}|...;k\rightarrow k+1)}{\rho_t(\alpha_{k+1}|...;k\rightarrow k+1)}\right\}$.
• Gibbs block to update $j^*_{k+1}$:
    Draw $j^*_{k+1}$ from $\varrho_t(j^*_{k+1}|...;k\rightarrow k+1)$.
```

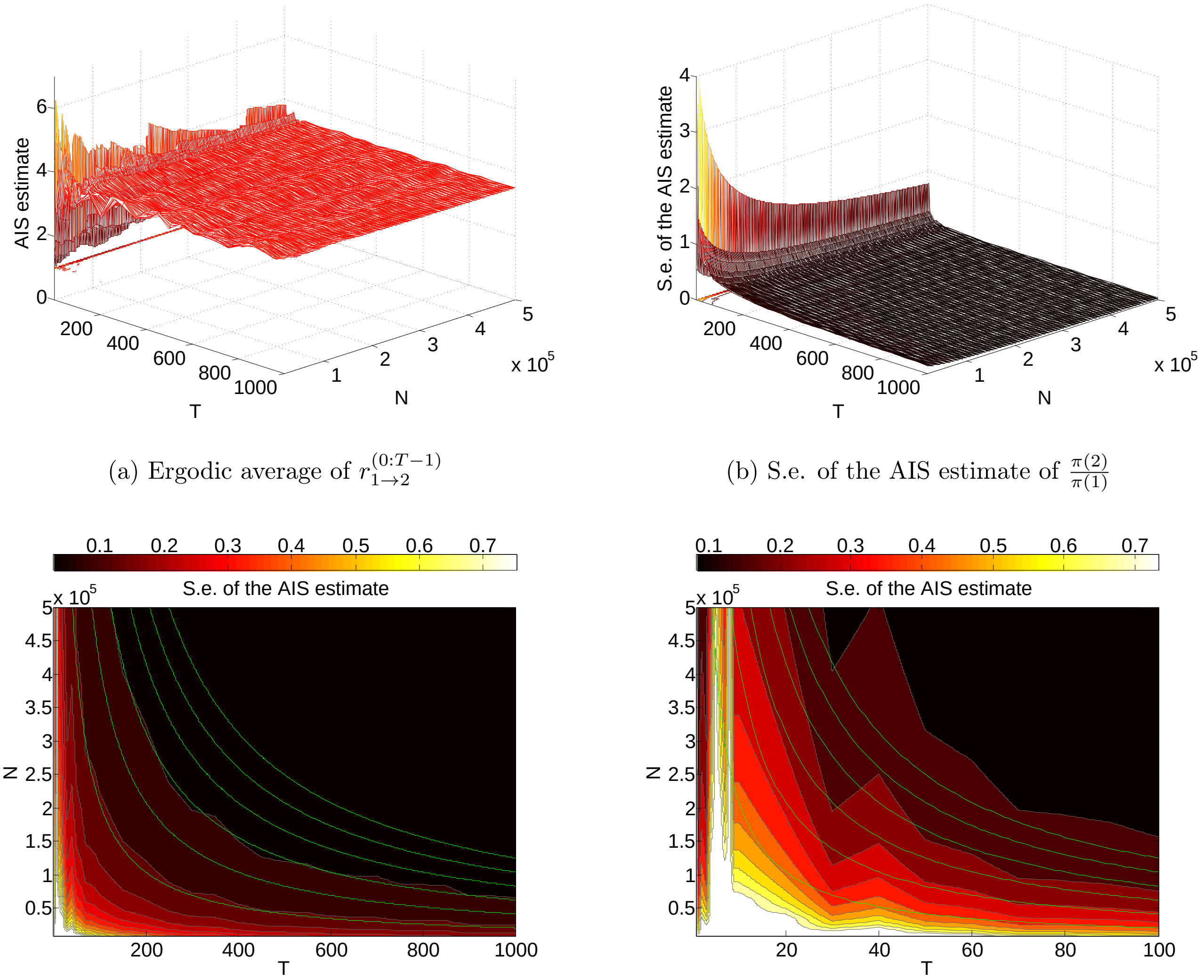
In order to assess the benefits of aisRJ we have numerically estimated the variance of estimators as a function of N and T for a wide range of values. In Fig. 5.6a and 5.7a, we report the ergodic averages of $r^{(0:T-1)}_{1\rightarrow 2}$ and $r^{(0:T-1)}_{2\rightarrow 3}$, which yield estimators of the ratios $\pi(2)/\pi(1)$ and $\pi(3)/\pi(2)$ respectively, for varying values of N and T. We observe the apparent instability of the estimates for small values of T (which includes the standard RJ-MCMC algorithm) even for large values of N, the number of RJ-MCMC sweeps, while the estimators seem to become consistently reliable as T increases, even for moderate values of T and N. In the remaining panes of Fig. 5.6 and 5.7 we display different views of the estimated standard errors for both estimators as a function of N and T and again observe the dramatic improvement brought by the aisRJ for relatively small values of T. Again while, as expected, for fixed values of T the performance improvement as a function of N follows a rate of $N^{-1/2}$, the improvement as a function of T for fixed values of N is initially very sharp and much faster than the rate $T^{-1/2}$. The superimposed green lines correspond to the curves $C = \sqrt{NT}$ for a range of values of C. We notice that these curves seem to match closely the isocontours of the standard error. Making the assumption that the computational cost of the algorithm is $O(N \times T)$ (which is realistic since we expect the cost of evaluating the posterior distributions to dominate), this suggest that beyond some value T_0 the standard error decreases as $1/\sqrt{N \times T}$ and that one is free to increase either N or T to achieve equivalent performance for a fixed computational budget. However, in this discussion, we have focused on the efficiency of the exploration of the models but one should also bear in mind that one may be interested in within model parameters. In such situations one would rather consider large values of N and base the inference on the samples generated by according to $\pi(n, \theta_n)$ directly, rather than increase T and attempt to reweight the intermediate sampled of the AIS procedure.

5.3 Gaussian mixture models

We consider the classical Bayesian univariate finite Gaussian mixture model with unknown number of components as proposed by Richardson and Green (1997) and analyze the enzyme dataset used in the paper. In the context of mixtures of normals, the likelihood of n observations $y_{1:n}$ is assumed to be of the form

$$\mathcal{L}_k\left(y_{1:n} | w_{1:k,k}, \mu_{1:k,k}, \sigma^2_{1:k,k}\right) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_{j,k} f\left(y_i | \mu_{j,k}, \sigma^2_{j,k}\right) \ , \tag{5.4}$$

where $k \in \mathcal{K}$ is the number of mixture components, $w_{j,k}$ is the weight of the j-th component, (such that $w_{j,k} \geq 0$ and $\sum_{j=1}^{k} w_{j,k} = 1$), and $f(y_i | \mu_{j,k}, \sigma^2_{j,k})$ is the normal distribution density with mean $\mu_{j,k} \in \mathbb{R}$ and variance $\sigma^2_{j,k} \in (0, +\infty)$, for $1 \leq j \leq k$. We use the prior model and the associated hyperparameters suggested in (Richardson and Green, 1997), which introduces the random hyperparameter β_k for the prior of the variances within model k. The within model parameter to be inferred is therefore $\theta_k := (w_{1:k,k}, \mu_{1:k,k}, \sigma^2_{1:k,k}, \beta_k)$ and is defined on the space $\Theta_k = [0,1]^{k-1} \times \mathbb{R}^k \times (0, +\infty)^k \times (0, +\infty)$. Following Richardson and Green (1997), we enforce the following *identifiability constraints* on the means, i.e. $\mu_{j,k} \leq \mu_{j+1,k}$ for $j = 1, ..., k-1$, in order to handle the standard label switching issue.



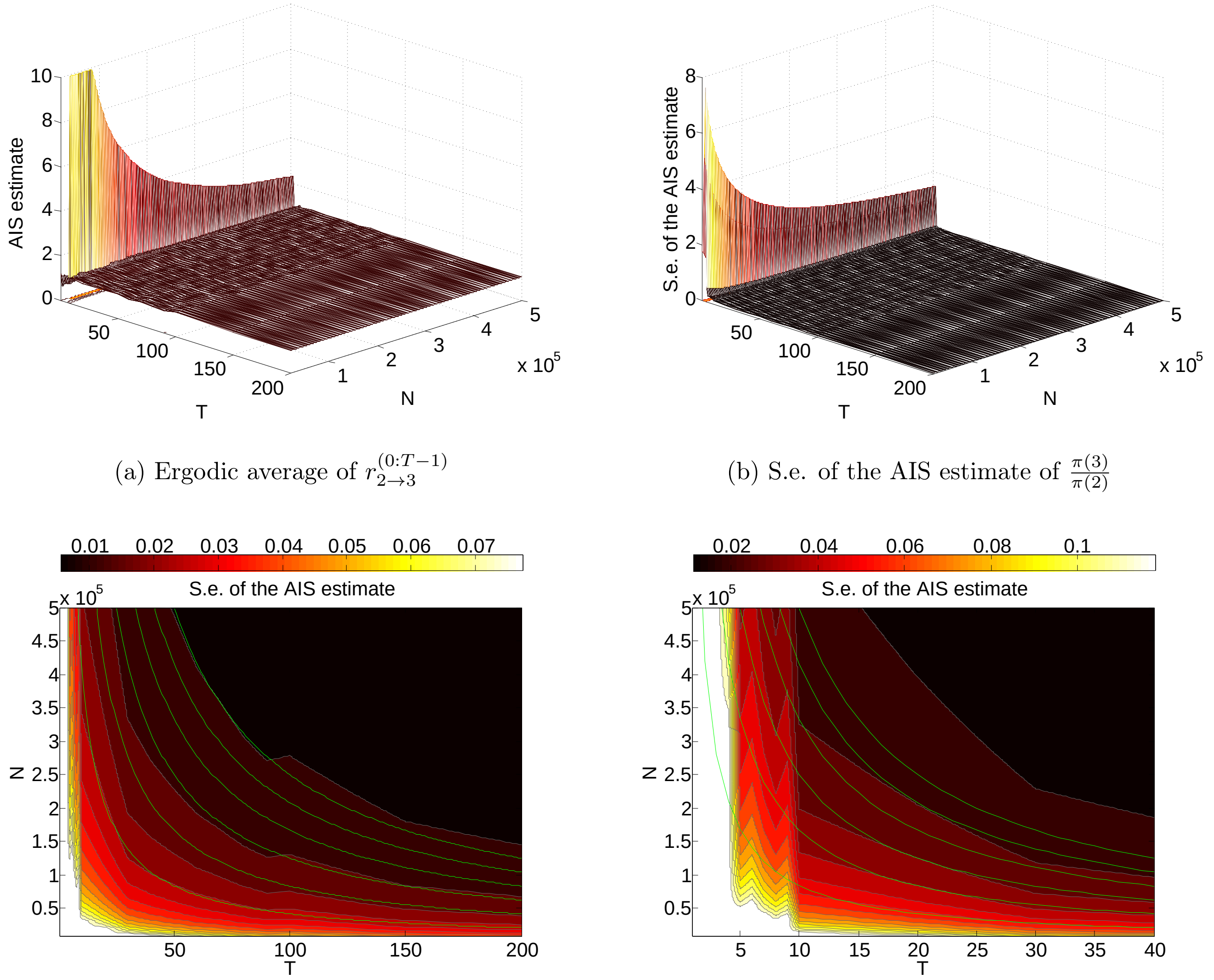
(a) Ergodic average of $r_{1\to 2}^{(0:T-1)}$

(b) S.e. of the AIS estimate of $\frac{\pi(2)}{\pi(1)}$

(c) S.e. of the AIS estimate of $\frac{\pi(2)}{\pi(1)}$

(d) S.e. of the AIS estimate of $\frac{\pi(2)}{\pi(1)}$ (zoom in)

Figure 5.6: Ergodic average of $r_{1\to 2}^{(0:T-1)}$ and Monte Carlo standard error of AIS estimate $\frac{\pi(2)}{\pi(1)}$. The mesh points are $T = \{1, ..., 1000\}$ and for $N = \{1, ..., 5 \cdot 10^5\}$. The contour plot has 40 levels. The curved green lines correspond to constant total annealing costs C.



(a) Ergodic average of $r_{2\to 3}^{(0:T-1)}$

(b) S.e. of the AIS estimate of $\frac{\pi(3)}{\pi(2)}$

(c) S.e. of the AIS estimate of $\frac{\pi(3)}{\pi(2)}$

(d) S.e. of the AIS estimate of $\frac{\pi(3)}{\pi(2)}$ (zoom in)

Figure 5.7: Ergodic average of $r_{2\to 3}^{(0:T-1)}$ and Monte Carlo standard error of AIS estimate $\frac{\pi(3)}{\pi(2)}$. The mesh points are $T = \{1, ..., 200\}$ and for $N = \{1, ..., 5 \cdot 10^5\}$. The contour plot has 40 levels. The curved green lines correspond to constant total annealing costs C.

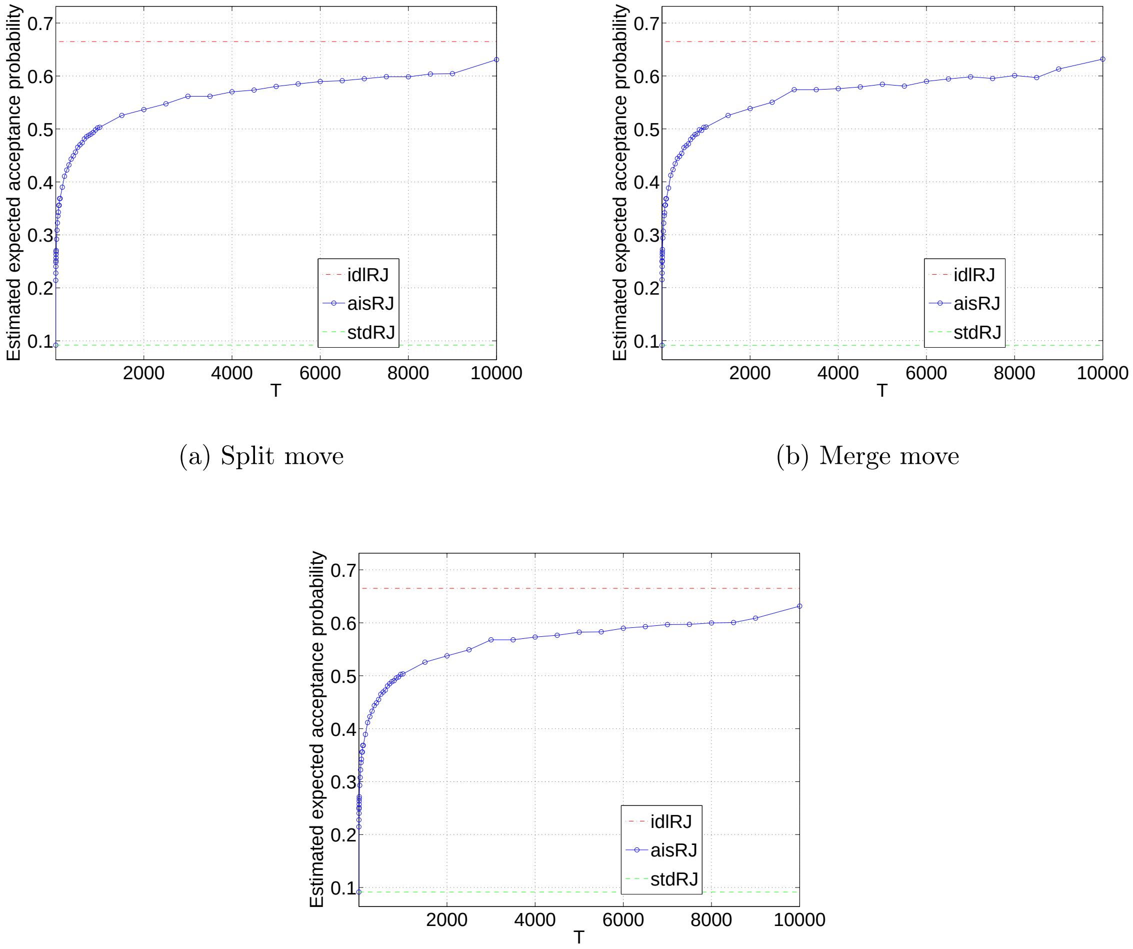
We have tested our aisRJ in the particularly challenging scenario where the local split/merge move of Richardson and Green (1997) was the only pair of moves used to communicate between models, and ignored their birth/death move. This pair of moves consists of either splitting one component into two neighboring ones or combining two adjacent components into one by using a moment matching strategy as described in (Richardson and Green, 1997). In order to describe our algorithm, it is sufficient here to introduce $u_{k+1\to k} = j^*_{k+1}$ (distributed according to a uniform distribution on $\{1, \ldots, k\}$), which allows one to choose the components j^*_{k+1} and $j^*_{k+1} + 1$ to be merged. We therefore have here that $\tilde{\theta}_{k+1} = (\theta_{k+1}, j^*_{k+1})$, which is defined on the extended space $\tilde{\varTheta}_{k+1} = \varTheta_{k+1} \times \{1, ..., k\}$. The random variable $u_{k\to k+1}$ and the corresponding $\varphi_{k\to k+1}$ are as in (Richardson and Green, 1997). We have tried both the geometric and arithmetic construction for the sequence of intermediate distributions, and have found the arithmetic approach to outperform the geometric construction. The results we present below correspond to the latter construction. As for the earlier example, we ensure the required reversibility conditions for $\{K_t,\ t = 1, \ldots, T\}$ and use a random permutation blockwise MCMC sweeps whose components are given in (ALG6).

In Fig. 5.8, we report the estimated expected acceptance probabilities for each move and their combination as T increases. Again the expected acceptance probabilities increase as the total annealing time T increases and converge to those of idlRJ. In Fig. 5.9-5.10, we report the autocorrelation function of k and $\mathbb{I}_{\{3\}}(k)$ as a function of T and again observe the same and now familiar phenomenon.

In Fig. 5.11 and 5.12, we report the Monte Carlo standard errors of the estimators of $\hat{\pi}(3)$ and $\hat{\pi}(4)$ as a function of N and T and observe again a sharp initial improvement in performance as T increases. As for the previous example we have superimposed the curves $C = \sqrt{N \times T}$ for various values of C and although their matching with the isocontours of the standard error is not as clear as for the earlier example, we again conclude that given a computational budget, it is preferable to increase N beyond some value T_0.

6 Conclusions

The implementation of efficient reversible jump MCMC (Green, 1995) algorithms, of interest for example in the context of model selection in a Bayesian framework, is notoriously difficult. In this paper, we show how it is possible to combine ideas from Jarzynski (1997b,a) and Neal (2004)

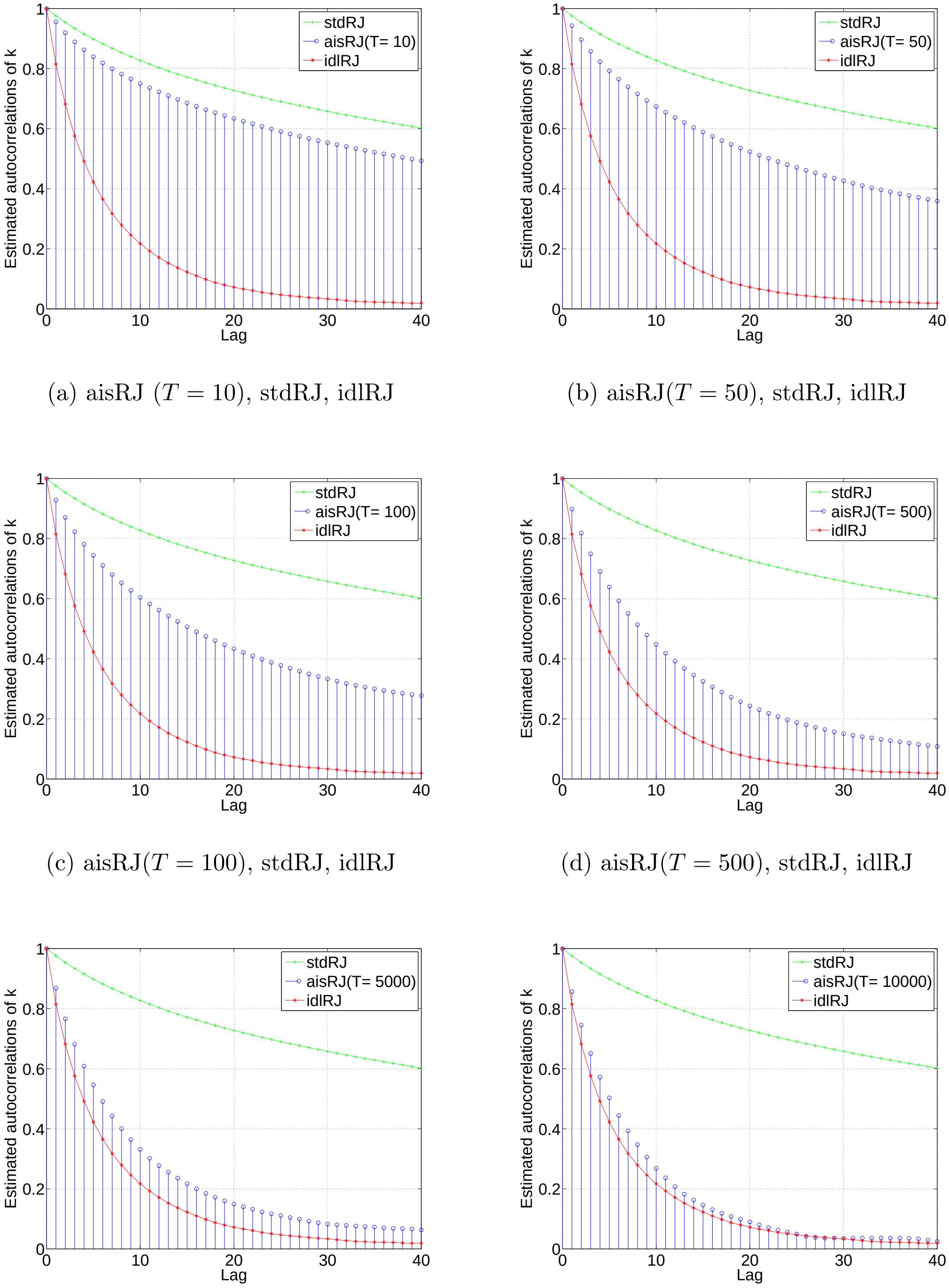


(a) Split move

(b) Merge move

(c) Split & Merge move

Figure 5.8: Estimated expectations of the acceptance probabilities for the complete split & merge pair of moves, the split move only and the merge move only. The involved algorithms in the plot are the stdRJ (- - -), the aisRJ (–o–) and the approximated idlRJ (-·-). The expected acceptance probabilities for stdRJ are 0.0910 for the Split move, 0.0920 for the Merge move and 0.0916 for the whole reversible pair of moves. The expected acceptance probabilities for the idlRJ is 0.6796 for a Split move, 0.6806 for a Merge move, and 0.6801 for the whole reversible pair of moves (i.e. the expectation of the latter two at stationarity).



(a) aisRJ ($T = 10$), stdRJ, idlRJ

(b) aisRJ($T = 50$), stdRJ, idlRJ

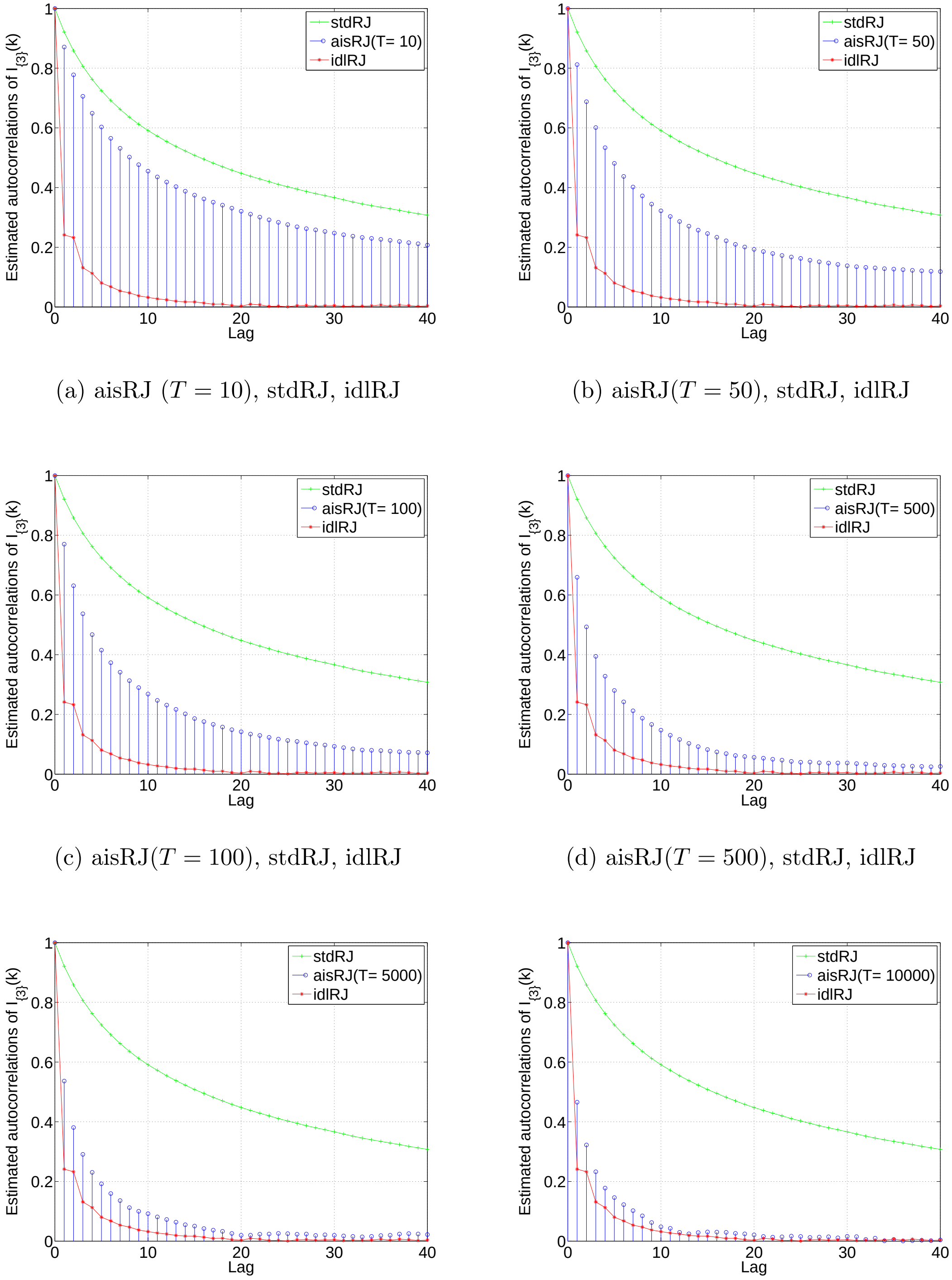
(c) aisRJ($T = 100$), stdRJ, idlRJ

(d) aisRJ($T = 500$), stdRJ, idlRJ

(e) aisRJ($T = 5000$), stdRJ, idlRJ

(f) aisRJ($T = 10000$), stdRJ, idlRJ

Figure 5.9: Autocorrelation function plot of k. The involved algorithms in the plot are the stdRJ (—), the aisRJ (–o) and the approximated idlRJ (—). The number of iterations N is equal to $2 \cdot 10^5$.



(a) aisRJ ($T = 10$), stdRJ, idlRJ

(b) aisRJ($T = 50$), stdRJ, idlRJ

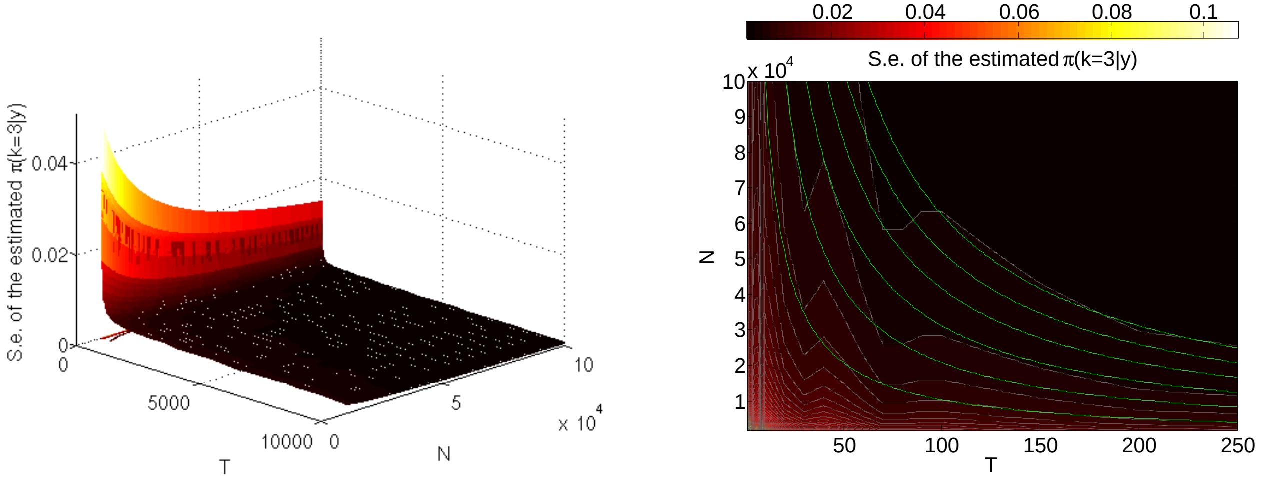
(c) aisRJ($T = 100$), stdRJ, idlRJ

(d) aisRJ($T = 500$), stdRJ, idlRJ

(e) aisRJ($T = 5000$), stdRJ, idlRJ

(f) aisRJ($T = 10000$), stdRJ, idlRJ

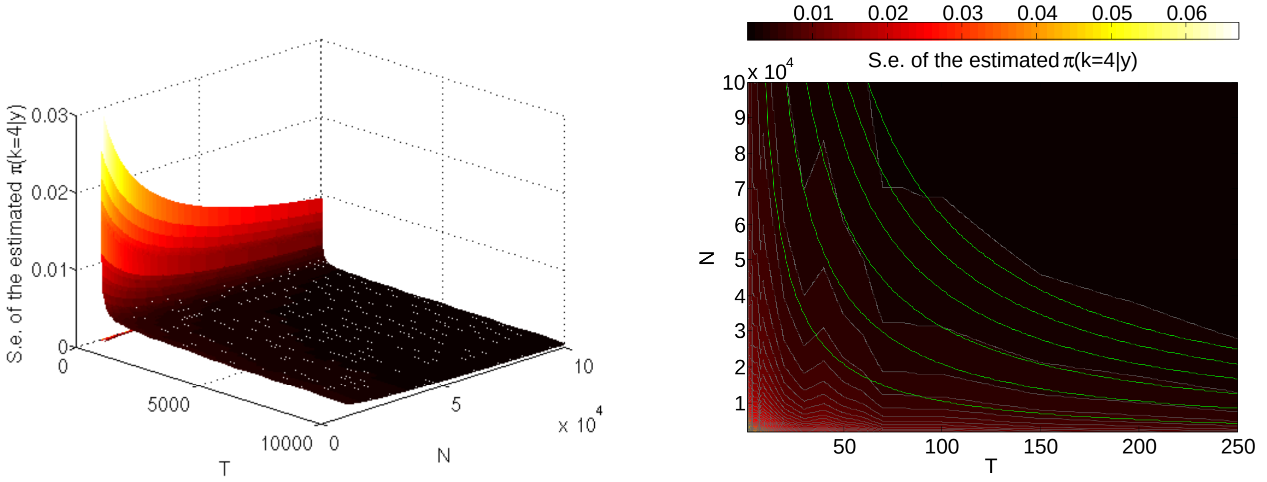
Figure 5.10: Autocorrelation function plot of $\mathbb{I}_{\{3\}}(k)$. The involved algorithms in the plot are the stdRJ (—), the aisRJ (−o) and the approximated idlRJ (—). The number of iterations N is equal to $2 \cdot 10^5$.



(a) Mesh plot

(b) Contour plot (zoom in)

Figure 5.11: Monte Carlo standard error of $\hat{\pi}(3)$ as a function of the total annealing time T and number of iterations N. The mesh points are $T = \{1, ..., 10000\}$ and for $N = \{1, ..., 10^5\}$. The contour plot has 40 levels. The curved green lines correspond to constant total annealing costs C.



(a) Mesh plot

(b) Contour plot (zoom in)

Figure 5.12: Monte Carlo standard error of $\hat{\pi}(4)$ as a function of the total annealing time T and number of iterations N. The mesh points are $T = \{1, ..., 10000\}$ and for $N = \{1, ..., 10^5\}$. The contour plot has 40 levels. The curved green lines correspond to constant total annealing costs C.

Algorithm 6 Components of the blockwise MCMC update.

```
• Random walk Metropolis in logit scale which updates the weights:
  Draw $\epsilon'_{w_{j,k+1}} \sim \mathcal{N}\left(0, \nu_{w_{j,k}}\right)$ and set $w'_{j,k}$ such that $\log\left(\frac{w'_{j,k+1}}{1-\sum_{i=1}^{k} w'_{i,k+1}}\right) = \log\left(\frac{w_{j,k+1}}{1-\sum_{i=1}^{k-1} w_{i,k+1}}\right) + \epsilon'_{w_{j,k+1}}$, for all $j = 1, ..., k$, and $w'_{k+1,k+1} = 1 - \sum_{j=1}^{k} w'_{j,k+1}$.
  Accept $w'_{1:k+1,k+1}$ with prob. $\min\left\{1, \frac{\rho_t(w'_{1:k+1,k+1}|...;k\to k+1)}{\rho_t(w_{1:k+1,k+1}|...;k\to k+1)} \prod_{j=1}^{k+1} \frac{w'_{j,k+1}}{w_{j,k+1}}\right\}$.
• Random walk Metropolis which updates the means:
  Draw $\epsilon'_{\mu_{j,k+1}} \sim \mathcal{N}\left(0, \nu_{\mu_{j,k+1}}\right)$ and set $\mu_{j,k+1}$ such that $\mu'_{j,k+1} = \mu_{j,k+1} + \epsilon'_{\mu_{j,k+1}}$, for all $j = 1, ..., k+1$.
  Accept $\mu'_{1:k+1,k+1}$ with prob. $\min\left\{1, \frac{\rho_t(\mu'_{1:k+1,k+1}|...;k\to k+1)}{\rho_t(\mu_{1:k+1,k+1}|...;k\to k+1)}\right\}$.
• Random walk Metropolis in log scale which updates the variances:
  Draw $\epsilon'_{\sigma^2_{j,k+1}} \sim \mathcal{N}\left(0, \nu_{\sigma_{j,k+1}}\right)$ and set $\sigma'^2_{j,k+1}$ such that $\log\left(\sigma'^2_{j,k+1}\right) = \log\left(\sigma^2_{j,k+1}\right) + \epsilon'_{\sigma^2_{j,k+1}}$, for all $j = 1, ..., k+1$.
  Accept $\sigma'^2_{1:k+1,k+1}$ with prob. $\min\left\{1, \frac{\rho_t(\sigma'^2_{1:k+1,k+1}|...;k\to k+1)}{\rho_t(\sigma^2_{1:k+1,k+1}|...;k\to k+1)} \prod_{j=1}^{k+1} \frac{\sigma'^2_{j,k+1}}{\sigma^2_{j,k+1}}\right\}$.
• Random walk Metropolis in log scale which updates $\beta_{k+1}$:
  Draw $\epsilon'_{\beta_{k+1}} \sim \mathcal{N}\left(0, \nu_{\beta}\right)$ and set $\beta_{k+1}$ such that $\log\left(\beta_{k+1}\right) = \log\left(\beta_{k+1}\right) + \epsilon'_{\beta_{k+1}}$.
  Accept $\beta_{k+1}$ with probability $\min\left\{1, \frac{\rho_t(\beta'_{k+1}|...;k\to k+1)}{\rho_t(\beta_{k+1}|...;k\to k+1)} \frac{\beta'_{j,k+1}}{\beta_{j,k+1}}\right\}$.
• The Gibbs block to update $j^*$:
  Draw $j^*$ from $\varrho_t(j^*_{k+1}|...;k\to k+1)$.
```

in order to facilitate the practical and efficient design of such algorithms. A crucial feature of the approach is that it allows one to approximate to an arbitrary degree an idealized Metropolis-Hastings algorithm with potentially very good convergence properties. This is achieved by adding T artificial bridging models between the models of interest in order to construct efficient transitions even in situations where a standard RJ-MCMC algorithm would not perform well. This naturally comes at an additional computational cost per model transition. However empirical evaluation on various models suggest that even moderate values of T can lead to dramatic improvements in the asymptotic error of estimators (at a rate superior to the standard Monte Carlo rate of $T^{-1/2}$) over standard RJ-MCMC implementations, making the approach very competitive when designing good reversible jump MCMC moves is difficult. There is however a trade-off between T and N the number of "outer loop" updates : our numerical experiments indicate the existence of an optimal value T_0 which is such that increasing T beyond this threshold or N leads to identical performance. This phenomenon is also mentioned in (Hendrix and Jarzynski, 2001). For practical purposes it is nevertheless important to be able to choose N as large as possible in order to allow for the reliable estimation of within model parameters. Given a fixed computational budget this therefore requires knowledge of T_0, which is not available in practice. Future work involves the theoretical determination of T_0 and the design of adaptive MCMC algorithms strategies (Andrieu and Thoms, 2008) that would allow one to determine such a value automatically.

ACKNOWLEDGMENTS

We would like to thank the three reviewers for their careful reading of an earlier version of the manuscript and their constructive comments.

SUPPLEMENTAL MATERIALS

Supplemental materials for the article are available online. All the supplemental files are included in a single archive. (aisRJ.code.tar.gz, GNU zipped tar file)

README file Description of the supplemental materials provided along with the paper.

(README, text file)

Code for Ex. 5.1 Python code used in the article (Ex. 5.1) for the demonstration of the aisRJ algorithm.

(aisRJ.toy.tar.gz, GNU zipped tar file)

Code for Ex. 5.2 FORTRAN code used in the article (Ex. 5.2) to implement aisRJ algorithm on the Poisson multiple change point model. The "coal mining disasters" dataset is included.

(aisRJ.CPT.tar.gz, GNU zipped tar file)

Code for Ex. 5.3 FORTRAN code used in the article (Ex. 5.3) to implement aisRJ algorithm on the Gaussian mixture model example. The "enzyme" dataset is included.

(aisRJ.Nmix.tar.gz, GNU zipped tar file)

References

Al-Awadhi, F., Hurn, M., and Jennison, C. (2004). Improving the acceptance rate of reversible jump MCMC proposals. *Statistics & probability letters*, 69(2):189–198.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

Andrieu, C. and Vihola, M. (2012). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *ArXiv e-prints*.

Brooks, S., Giudici, P., and Roberts, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–55.

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

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Hendrix, D. and Jarzynski, C. (2001). A “fast growth” method of computing free energy differences. *The Journal of Chemical Physics*, 114(14):5974–5981.

Jarzynski, C. (1997a). Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E*, 56:5018–5035.

Jarzynski, C. (1997b). Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693.

Karagiannis, G. (2011). *Annealed Importance Sampling Reversible Jump Markov chain Monte Carlo*. PhD thesis, University of Bristol, UK.

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

Neal, R. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Neal, R. (2004). Taking bigger Metropolis steps by dragging fast variables. Technical Report 0411, Dept. of Statistics, University of Toronto.

Richardson, S. and Green, P. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Roberts, G. and Rosenthal, J. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.