Tempered Bayesian Algorithms

Ruth King

University of St. Andrews

This is joint work with Bobby Gramacy and Richard Samworth at the University of Cambridge

Tempered Bayesian Algorithms OR Importance (Sampling applied to Simulated) Tempering

Ruth King

University of St. Andrews

This is joint work with Bobby Gramacy and Richard Samworth at the University of Cambridge

Overview

- Introduction
- (Simple) Motivating Example
- Simulated Tempering
- Importance Sampling

} *Importance Tempering*

- Example

  (and application to RJMCMC)

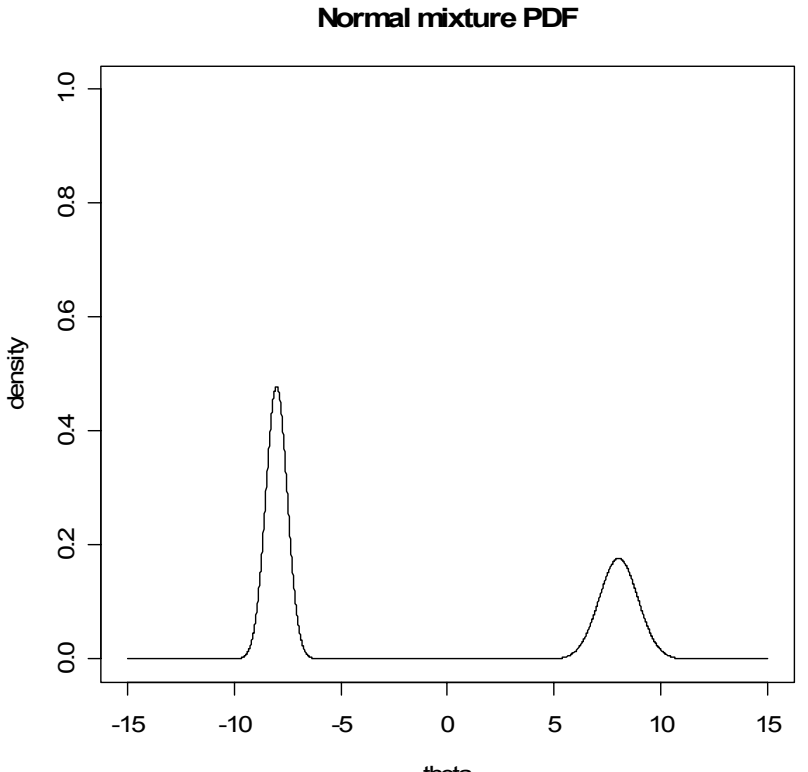
Introduction

- We will address the issue of poor mixing within (RJ)MCMC algorithms.
- In particular, we consider the case where the distribution is potentially multi-modal.
- The methods are easily extended to the case of model uncertainty and the use of the RJMCMC algorithm.
- We will demonstrate the potential increase in efficiency that can be obtained for little additional computational expense.

(Really) Simple Motivating Example

- Suppose that the distribution we wish to sample from is a mixture of two Normal distributions:

$$0.6\ N(-8, 0.5^2) + 0.4\ N(8, 0.9^2)$$



Simple Motivating Example

- Performing a random walk MH algorithm will typically mean that only mode will be explored within the MCMC iterations.
- However, using a random walk with large variance will typically result in very poor mixing with a large rejection probability.
- Pilot-tuning can be used to identify the individual modes (starting the chain at different points).
- A MH algorithm can then be tuned to allow movements between modes.
- Note that this requires the modes to be identified via pilot-tuning……

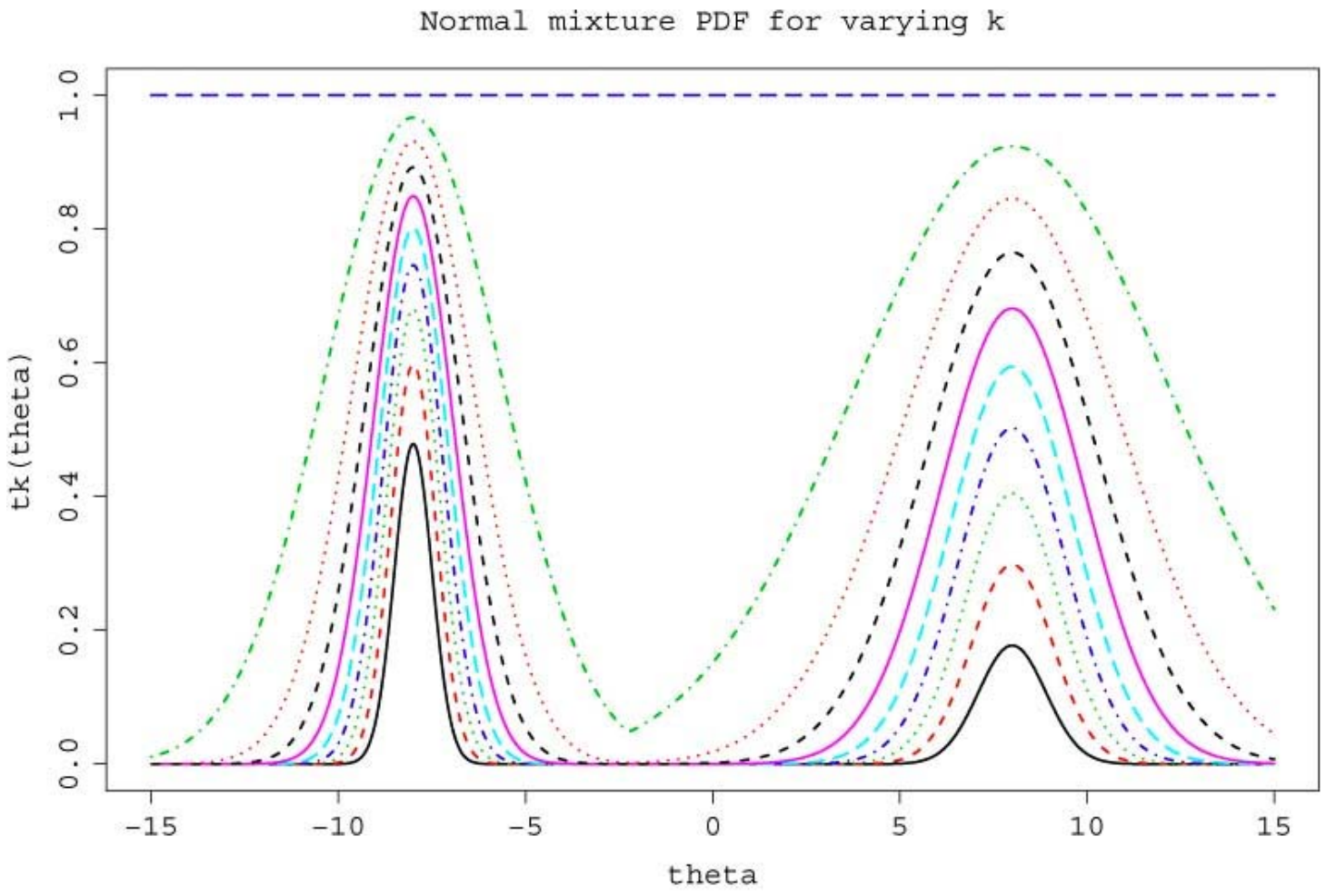
Simulated Tempering

- Suppose that we are interested in sampling from the distribution $\pi(\theta)$.
- Introduce auxiliary variable, k, and define the joint distribution,

$$\pi(\theta,k) = \pi(\theta)^k \, p(k),$$

  k is typically called the (inverse) temperature.
- When $k=1$, $\pi(\theta,k) \propto \pi(\theta)$.
- When $k<1$, the temperature is raised, and the distribution is "flattened" or "squashed".
- For $k=0$, $\pi(\theta,k) \propto 1$.

Simulated Tempering



Simulated Tempering

- To obtain a sample from $\pi(\theta)$:
  - Set a temperature ladder, (i.e. the set of possible values for $k = k_1,\ldots,k_m$). We consider a geometric ladder:
    $$k_i = (1+\Delta_k)^{1-i} \quad \text{for } i=1,\ldots,m.$$
  - Sample from the joint distribution, $\pi(\theta,k)$, using standard MCMC updates (MH/Gibbs for θ and MH for k – propose to move to neighbouring k with equal probability, i.e. for k_i, $P(k_i \rightarrow k_{i+1}) = P(k_i \rightarrow k_{i-1}) = 0.5$).
  - Obtain posterior estimates of $\pi(\theta)$ by retaining the sampled values of θ when $k=1$.
- Typically, the pseudo-prior $p(k)$ is set such that the chain spends (approximately) equal amounts of time in each temperature.

Simulated Tempering/Importance Sampling

- Simulated tempering can be regarded as wasteful in that only values of θ are used, when $k = 1$.
- However, there is information in the other θ values sampled when $k \neq 1$ regarding the distribution $\pi(\theta)$.
- This is where the idea of importance sampling comes in – we can reweight the values of θ for the other temperatures ($k \neq 1$) to obtain summary estimates of $\pi(\theta)$ using all the sampled values.
- (Note this is not a new idea!).

(Naïve) Importance Sampling

- Suppose that we wish to estimate $E_{\pi}(h(\theta))$.
- We can use importance sampling to estimate this by,

$$h^*_{IS} = W^{-1} \sum^{T}_{t=1} w(\theta^t, k^t) h(\theta^t)$$

$$w(\theta, k) = \pi(\theta)/\pi(\theta)^k$$
$$W = \sum^{T}_{t=1} w(\theta^t, k^t)$$

- We now describe how we can improve on this estimator (and will demonstrate how poor this naïve IS estimator can be).

Importance Tempering

- Note that we can obtain an importance sampling estimate of $E_\pi(h(\theta))$ for each temperature, k_i, which we denote by h_i^*.
- We consider an estimator of $E_\pi(h(\theta))$ of the form,

$$h^* = \sum \lambda_i h_i^*$$

  where $0 \leq \lambda_i \leq \sum \lambda_i = 1$.
- Note – naïve IS and ST are both special cases of this general algorithm.
- We find an "optimal" set of values for λ_i.
- We define optimal in terms of maximising the effective sample size.

Effective Sample Size (ESS)

- Following Liu (2001) we define the ESS as

$$\text{ESS} = \frac{T}{1 + \text{cv}^2}$$

where,

$$\text{cv}^2 = \frac{\sum_{t=1}^{T} (w(\theta^t, k^t) - \bar{w})^2}{(T\text{-}1)\,\bar{w}^2}$$

- This can be regarded as a measure of efficiency of the given IS algorithm.

Optimal choice of λ

- Recall that $h^* = \sum \lambda_i h_i^*$.
- The value of the λ_i's that maximises the ESS is given by,

$$\lambda_i^* = \frac{\beta_i}{\sum^m_{i=1} \beta_i}$$

where

$$\beta_i = \frac{W_i^2}{\sum^{T_i}_{j=1} w_{ij}^2}$$

such that w_{ij} denotes the weight of the j^{th} realisation in temperature k_i; and W_i the sum of the weights for temperature k_i.

Examples – toy example

- We return to the toy example of the mixture of two Normal distributions.
- We run the MCMC iterations for 100000 iterations and compare the ESS for the naïve IS, ST and IT approaches (m=40; k_m = 0.1)

Method	ESS
ST	2535
Naïve IS	17779
IT	22913

Examples - RJMCMC

- We now apply IT to an example where there is model uncertainty.
- We consider mark-recapture-recovery data of shags on the Isle of May (Scotland), where there are three "sets" of parameters:
  - $\phi_{a,t}$ – survival probability at time t for individual aged $a = \{1,2,3,A\}$
  - $\lambda_{a,t}$ – recovery probability at time t for individual aged $a = \{1,2,3,A\}$
  - $p_{a,t}$ – recapture probability at time t for individual aged $a = \{1,2,3,A\}$.

Models

- There are a large number of possible models for ϕ, λ and p, corresponding to age and/or time dependence.
- Typically we can denote the models in the form:

  $$\phi_1(t),\ \phi_{2,3}(t),\ \phi_A \,/\, p_{1,2,3,A}(t) \,/\, \lambda_1(t),\ \lambda_{2,3},\ \lambda_A$$

  where the subscripts denote the age dependence; and the (t) corresponds to time dependence for the given parameters.
- Clearly there are a number of possible models for each set of parameters.
- A RW MH is used for each parameter, conditional on the model, and appears to perform well.

RJMCMC

- Moving between the different possible models for each set of parameters is difficult.
- This is largely as a result of the large difference in the number of parameters between "neighbouring" models.
- For example, adding/removing time dependence, means changing the dimension of the model by 8 parameters.
- Alternatively, adding/removing age dependence results in a difference of 1 or 9 parameters (dependent on whether the parameter(s) are age time dependent or not).

RJMCMC

- In order to move between the different models we perform an initial pilot run in the saturated model (i.e. fully age and time dependent).
- When proposing to move between different models, we set the mean of the proposal distribution of the parameters to be a function of the posterior mean of the parameters from the saturated model.
- Eg suppose we propose to move from a model with $p_1(t)$, $p_2(t)$ to the model with $p_{1,2}(t)$. We propose,

$$p_{1,2}(t) \sim N(\mu(t),\sigma^2),$$

with,

$$\mu(t) = 0.5(\mu_1(t)+\mu_2(t))$$

Improving the RJMCMC algorithm

- With extensive pilot-tuning (including different proposal distributions), the acceptance probabilities for moving between different models are still small.
- This means that movement between non-neighbouring models is very difficult.
- Even with starting from over-dispersed starting values, we may not spot "multi-modality" over the model space.
- We implement the IT algorithm to (hopefully) improve the mixing between the different models.
- We set m=40 and k_m = 0.1.

IT Results - ESS

- Simulations are run for 10^7 iterations with the initial 10% discarded as burn-in.

Method	ESS
ST	248158
Naïve IS	5
IT	612026

- The catastrophic ESS for the naïve IS is a result of a few very large weights obtained at hot temperatures (i.e. for small k).

IT Results – Acceptance probs

- As previously discussed, moving between different models can be difficult.
- Thus, we now compare the acceptance probabilities for the standard RJMCMC algorithm, and corresponding IT algorithm.

	Mean % acceptance rate			
Method	Split age	Merge age	Add time	Remove time
RJMCMC	1.30	0.50	0.01	0.14
IT	1.32	1.21	0.30	1.45

IT Results – Models visited

- We can also compare the number of different models visited for the different methods.

	ϕ				λ				p			
	max	IT	ST	RJ	max	IT	ST	RJ	max	IT	ST	RJ
Age+time	54	51	26	18	94	12	5	3	94	75	25	28
Age	10	7	4	3	15	7	4	2	15	15	11	12

- Overall total number of models visited for each method was 3080 for IT; 177 for ST and 233 for standard RJMCMC.
- Posterior estimates (e.g. posterior means, model probabilities) were very similar for RJMCMC & IT)

Summary/comments

- Naïve importance sampling can lead to very poor ESS and corresponding estimates
- IT can be implemented with minimal additional computational (and programming) effort post-process.
- Comparing IT results with standard RJMCMC results can reassure us (but not guarantee) that we have not missed models with high posterior mass.

Future work/improvements

- Typically, within ST the MH updates are adaptive to the temperature – we have not applied such a method to the RJMCMC updates in our example.
- It is possible to consider alternative temperature ladders to the geometric ladder – is it possible to find some form of "optimal" ladder? Or consider continuous values for k?
- The ESS considered does not include any autocorrelation between successive draws from the Markov chain – extend the idea of ESS to include serial autocorrelation.