Resampling from the past to improve on MCMC algorithms[1]

Yves F. Atchadé[2]

(May 2006)

Abstract

We introduce the idea that resampling from past observations in a Markov Chain Monte Carlo sampler can fasten convergence. We prove that proper resampling from the past does not disturb the limit distribution of the algorithm. We illustrate the method with two examples. The first on a Bayesian analysis of stochastic volatility models and the other on Bayesian phylogeny reconstruction.

Key words: Monte Carlo methods, Resampling, Stochastic volatility models, Bayesian phylogeny reconstruction

MSC Numbers: 60C05, 60J27, 60J35, 65C40

1 Introduction

Markov Chain Monte Carlo (MCMC) methods have become the standard computational tool for bayesian inference. But the great flexibility of the method comes with a price. Namely, it is very difficult to determine a priori (before the simulation) or a posteriori whether a given MCMC sampler can mix or has mixed in a given computing time. The challenge becomes that of designing fast converging Monte Carlo algorithms. Contributions in this field can have significant impact in other scientific disciplines where these methods are used.

[1]This work is funded in part by NSERC Canada

[2]Department of Mathematics and Statistics, University of Ottawa, email: yatchade@uottawa.ca

In this paper, we propose a new and general approach to increase the convergence rate of MCMC algorithms. The method is based on resampling. Suppose that at time n, we want to sample X_n in a MCMC algorithm. Instead of sampling X_n from $P(X_{n-1}, \cdot)$ for some transition kernel P, we propose to obtain X_n by resampling independently from $\{X_B, \ldots, X_{n-1}\}$, where $B \geq 0$ is some burn-in period. This *resampling from the past* step is then repeated during the simulation at some predetermined times $a_1 < a_2 < \ldots$. Basically, the idea is to look at $\{X_B, \ldots, X_{n-1}\}$ as a sample from π. Therefore resampling from the past allows the sampler to move more easily and according to a distribution that is close to π. The resampling schedule plays an important role. As long as we do not resample too much (typically, we need (a_n) such that $a_n/n \to \infty$ as $n \to \infty$), we show that resampling from the past does not disturb the limit distribution of the sampler.

Resampling from the past can perform poorly if the original sampler has a very poor convergence rate. We extend the framework above by allowing resampling from an auxiliary process $\{X_n^{(0)}\}$ that has a better convergence rate towards its target distribution $\pi^{(0)}$. Resampling from an auxiliary process is not new and is the idea behind the equi-energy sampler recently proposed by (Kou et al., 2006). But the equi-energy sampler has a number of complications that we avoid here by using an *importance-resampling*. The idea is also apparent in the "Metropolis with an adaptive proposal" of (Chauveau and Vandekerkhove, 2001). On the theoretical side, we show in the case of *importance-resampling*, that resampling from an auxiliary process does not disturb the limit distribution of the sampler.

We apply our methods to two examples from Bayesian data analysis. First, we

consider the Bayesian analysis of stochastic volatility models (Kim et al., 1998). We improve the efficiency of the basic Gibbs sampler for this problem by a factor of fifty (50). In the second example, we look at Bayesian phylogenetic trees reconstruction. Our methods improve the efficiency of the MCMC sampler of (Larget and Simon, 1999) by a factor of hundred (100).

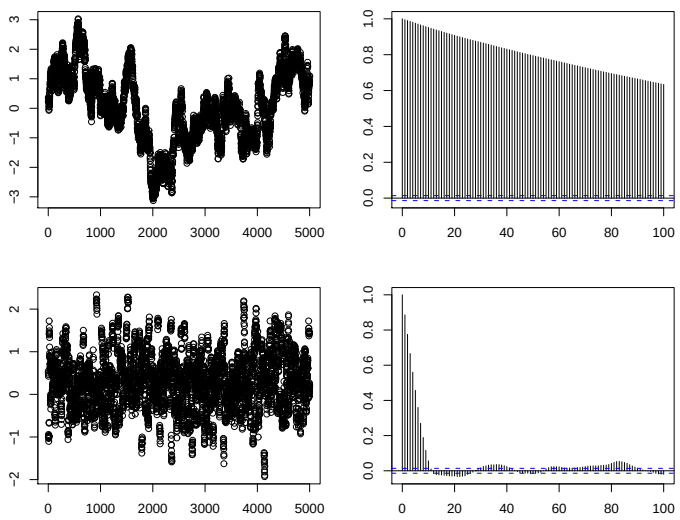
The paper is organized as follows. In Section 2, we present the idea of resampling from the past. Resampling from an auxiliary process is discussed in Section 3. All the theoretical proofs are postponed to Section 5 and the simulation examples are presented in Section 4.

2 Resampling from the past

Let $\{X_n\}$ be a Markov chain with state space $(\mathcal{X}, \mathcal{B})$, transition kernel P and invariant distribution π started at $X_0 = x$. If the chain is ergodic then $\mathcal{L}_x(X_n)$, the distribution of X_n, will converge to π as $n \to \infty$. But it is well known that for MCMC algorithms, the convergence of $\mathcal{L}_x(X_n)$ to π can be too slow for the sampler to be useful. We propose the following idea to accelerate the convergence of Markov chains. Suppose that after a burn-in period B, we have the sample $\{X_B, X_{B+1}, \ldots, X_{n-1}\}$ at time n. Instead of sampling $X_n \sim P(X_{n-1}, \cdot)$ as we normally do, we obtain X_n by resampling independently and with equal weight from $\{X_B, X_{B+1}, \ldots, X_{n-1}\}$. The resampling step is then repeated at some predetermined times $a_1 < a_2 < \ldots$. Intuitively, if P mixes reasonably well, $\{X_B, X_{B+1}, \ldots, X_{n-1}\}$ can be seen as a sample points from π and resampling will operate as an i.i.d. sampling from π.

Consider the following toy example. We want to use the Random Walk Metropolis (RWM) algorithm with proposal density $q(x,y) = \mathcal{N}(y-x;0,\sigma^2)$ with $\sigma = 0.1$ to sample from the standard normal density $\mathcal{N}(x;0,1)$; where $\mathcal{N}(x;\mu,\sigma^2)$ denotes the density of the normal distribution $N(\mu,\sigma^2)$ with mean μ and variance σ^2. We compare the plain RWM with a RWM with resampling. Each sampler is run for $25,000$ iterations. Graph 1 (a) shows the last $5,000$ sample points and Graph (b), the autocorrelation function from the last $20,000$ points in the plain RWM sampler. For the RWM with resampling, we resample at times $B + \lceil k^\alpha \rceil$ (see the justification below), with $B = 5,000$ and $\alpha = 1.3$. Graph 1 (c) and (d) show the corresponding results for the RWM with resampling. As we can see, there is a significant gain in efficiency.

Intuitively, resampling helps to the extend that P mixes rapidly. Differently put, the slower P converges to π, the longer we should wait between two resampling. What should be the resampling schedule (a_k)? Obviously, we should not resample all the time. We find that the choice $a_k = b_1 + b_2 k^\alpha$, $\alpha > 1$ is a valid choice and works well in practice for $b_2 = 1$, and $\alpha \approx 1.3$. The choice $a_k = b_1 + b_2 k$ is also theoretically valid as long as b_2, the time between two resampling, is large enough.



Graph 1: Comparing a plain RWM and a RWM with resampling in sampling from the standard norma distribution $N(0,1)$.

2.1 Theoretical discussion

What can we prove about this algorithm? We can prove that despite the resampling, the limit distribution of the algorithm is π under certain conditions on P and on the resampling schedule (a_k). We recall the algorithm. The resampling schedule $0 < a_1 < a_2 < \cdots < a_n < \infty$ is given and is nonrandom. Fix B the burn-in period. We start the sampler at some arbitrary point $X_0 = x$. At time $n \geq 1$, given $\{X_0, \ldots, X_{n-1}\}$, if $n > B$ and $n = a_k$ for some $k \geq 1$ then $X_n \sim \frac{1}{n-B} \sum_{j=B}^{n-1} \delta_{X_j}(\cdot)$. Otherwise sample $X_n \sim P(X_{n-1}, \cdot)$. We denote Pr the underlying probablity measure and $\mathbb{E}$ its expectation operator. Here are some standard notations that we use below. If P_1 and P_2 are two transition kernels on $\mathcal{X}$, the product P_1P_2 denotes the transition kernel $P_1P_2(x, A) := \int P_1(x, dy)P_2(y, A)$. Recursively, we can define P_1^n by $P_1^1 = P_1$

and $P_1^n = P_1^{n-1} P_1$. A transition kernel P_1 defines a linear operator (also denoted P_1) on the space of $\mathbb{R}$-valued functions on $(\mathcal{X}, \mathcal{B})$ into itself, by $P_1 f(x) := \int P_1(x, dy) f(y)$. If μ is a signed measure on $(\mathcal{X}, \mathcal{B})$, we denote $\mu(f) := \int \mu(dx) f(x)$ and we will also write μ to denote the linear functional on the space of $\mathbb{R}$-valued functions on $(\mathcal{X}, \mathcal{B})$ thus induced. Finally, we define $\mu P_1(A) := \int \mu(dx) P_1(x, A)$. Let $V : \mathcal{X} \to [1, \infty)$ be given. For $f : \mathcal{X} \to \mathbb{R}$, we define its V-norm $|f|_V := \sup_{x \in \mathcal{X}} \frac{|f(x)|}{V(x)}$ and we introduce the space $L_V := \{f : \mathcal{X} \to \mathbb{R} : |f|_V < \infty\}$. For a signed measure μ on $(\mathcal{X}, \mathcal{B})$ we define its V-norm $\|\mu\|_V := \sup_{f \in L_V, |f|_V \leq 1} |\mu(f)|$. Similarly, for a linear operator T from the space of $\mathbb{R}$-valued functions on $\mathcal{X}$ into itself, we define $\|\!|T|\!\|_V := \sup_{f \in L_V, |f|_V \leq 1} |Tf|_V$. If $\|\!|T|\!\|_V < \infty$, then T defines a bounded linear operator from the Banach space $(L_V, |\cdot|_V)$ into itself.

We assume that the transition kernel P in the algorithm is geometrically ergodic in the sense that:

Assumption (A): *P is irreducible, aperiodic and there exists $\rho \in (0, 1)$, a measurable function $V : \mathcal{X} \longrightarrow [1, \infty)$ such that*

$$\|\!|P^n - \pi|\!\|_V = O(\rho^n), \tag{1}$$

This assumption implies that $\pi(V) < \infty$ and that $\sup_n P^n V^\alpha(x) < \infty$ for any $x \in \mathcal{X}$, $\alpha \in [0, 1]$. We refer the reader to (Meyn and Tweedie, 1993) for more on geometrically ergodic Markov chains. This is a convenient assumption that is known to hold for many MCMC sampler.

Define $c := \frac{1}{1-\rho}$ and $\delta_n := -a_1 \log(\rho) + \sum_{k=2}^n \log(a_k) - \log(c + a_{k-1})$.

Theorem 2.1. *Assume (A). Then there exists a constant $C \in (0, \infty)$ such that for $a_k \leq n < a_{k+1}$:*

$$\left\| \mathcal{L}^{(n)} - \pi \right\|_V \leq C \rho^{n-k} \exp\left[-\delta_k\right], \tag{2}$$

where the transition kernel $\mathcal{L}^{(n)}$ is defined by $\mathcal{L}^{(n)}(x, A) := \Pr\left[X_n \in A | X_0 = x\right]$. In particular if $\delta_n \to \infty$ as $n \to \infty$, the algorithm has limit distribution π.

Proof. See Section (5). □

Resampling from the past can sensibly reduce the autocorrelation in the output of a MCMC algorithm. But when the sampler has a very slow mixing time, it might be better to resample from an auxiliary process that has a better mixing time.

3 Resampling from an auxiliary process

As above, $\pi(dx) \propto h(x)\lambda(dx)$ is the probability measure of interest on the measure space $(\mathcal{X}, \mathcal{B}, \lambda)$. We introduce another probability measure $\pi^{(0)}(dx) \propto h^{(0)}(x)\lambda(dx)$ on $(\mathcal{X}, \mathcal{B}, \lambda)$. Let $\{X_n^{(0)}\}$ be a Markov chain with invariant distribution $\pi^{(0)}$ and transition kernel $P^{(0)}$. Let $k: \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a measurable function and T a transition kernel on $(\mathcal{X}, \mathcal{B})$. Define the transition kernel $Q(x, dy) = \frac{\int \pi^{(0)}(dz) k(x,z) T(z,dy)}{\int \pi^{(0)}(dz) k(x,z)}$. Following (Tierney, 1998), let $S \subseteq \mathcal{X} \times \mathcal{X}$ be such that the probability measures $\pi(dx)Q(x, dy)$ and $\pi(dy)Q(y, dx)$ are mutually absolutely continuous on S and mutually singular on $\mathcal{X} \setminus S$.

We assume that $\{X_n^{(0)}\}$ converges (reasonably quickly) to $\pi^{(0)}$. Let P be a transition kernel with invariant distribution π and $\theta \in [0, 1]$. The algorithm works as

follows. Given $(X_0^{(0)}, \ldots, X_n^{(0)}, X_0, \ldots, X_n)$:

- with probability θ, we sample X_{n+1} from $P(X_n, \cdot)$;
- with probability $1-\theta$, we propose Y from $R_n(X_n, \cdot)$ where $R_n(x, A) = \frac{\sum_{l=0}^n k(x, X_l^{(0)}) T(X_l^{(0)}, A)}{\sum_{l=0}^n k(x, X_l^{(0)})}$. In other words, we resample Y_1 from $\{X_0^{(0)}, \ldots, X_n^{(0)}\}$ with weights $k(X_n, X_l^{(0)})$ and propose $Y \sim T(Y_1, \cdot)$.

  Then we either "accept" Y and set $X_{n+1} = Y$ with probability $\alpha(X_n, Y)$, or "reject" Y and set $X_{n+1} = X_n$ with probability $1 - \alpha(X_n, Y)$, where

$$\alpha(x, y) = \begin{cases} \min\left[1, \frac{\pi(dy)Q(y,dx)}{\pi(dx)Q(x,dy)}\right] & \text{if } (x, y) \in S \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

For n large enough, a sample from $R_n(x, \cdot)$ can be seen as a sample from $Q(x, dy)$ which explain the acceptance probability (3). But the algorithm is not feasible as such because the ratio in (3) cannot be computed in general. The natural choice which simplifies Q is to choose a transition kernel T that is invariant under π and $k(x, y) = \omega(y) = h(y)/h^{(0)}(y)$. With this choice, we get $\alpha(x, y) \equiv 1$ on S. We call this scheme *importance-sampling* resampling. It is not necessary to choose a complicated transition kernel for T. Throughout, we choose T to be the identity transition kernel, $T(x, A) = \mathbf{1}_A(x)$ in which case $S = \{(x, y) : \ 0 < h(x)k(x, y) < \infty\}$.

Another choice for which the acceptance ratio $\alpha(x, y)$ simplifies is $T(x, A) = \mathbf{1}_A(x)$ and $k(x, y) = \mathbf{1}_{\{D(x)\}}(y)$ where (D_i) is a given partition of $\mathcal{X}$ and $D(x) = D_i$ if $x \in D_i$. This corresponds to the set-up of the equi-energy sampler of (Kou et al., 2006). With this choice of k, the acceptance probability becomes $\alpha(x, y) = \min\left(1, \frac{\omega(y)}{\omega(x)}\right)$ (and 0 if $\omega(x) = 0$ or $\omega(x) = \infty$). The drawback with this choice is that we have to define

the partition (D_i) in the first place and an inadequate partition can result in a high rejection rate for the resampling step.

Algorithm 3.1 (MCMC with Importance-Resampling from an auxiliary process). *At some time $n \geq 1$, given $\left(X_0^{(0)}, \ldots, X_n^{(0)}, X_0, \ldots, X_n\right)$:*

(i) *With probability θ, sample X_{n+1} from $P(X_n, \cdot)$. Otherwise with probability $1-\theta$ sample X_{n+1} from*

$$\frac{\sum_{i=0}^{n} \omega(X_i^{(0)}) \delta_{X_i^{(0)}}(\cdot)}{\sum_{i=0}^{n} \omega(X_i^{(0)})}.$$

(ii) *Sample $X_{n+1}^{(0)}$ from $P^{(0)}(X_n^{(0)}, \cdot)$.*

3.1 Theoretical discussion

We look more closely to $\{X_n\}$ when the importance-resampling scheme is used. (Atchade and Liu, 2006) have shown that the limit distribution of the equi-energy sampler is indeed π under a number of conditions. We can study the process $\{X_n\}$ along the same line. The assumption we impose are less stronger than in (Atchade and Liu, 2006). We continue with the notations in Section 2.1. Essentially we will assume that $P^{(0)}$ is geometrically ergodic and that the weight function satisfies $|\omega|_{V^\alpha} < \infty$, for some $\alpha \in [0, 1/4)$. Typically ω is bounded.

Assumption (A0): *$P^{(0)}$ is irreducible and aperiodic and there exists $\rho_0 \in (0,1)$ such that*

$$\left\| P^{(0)^n} - \pi^{(0)} \right\|_V = O\left(\rho_0^n\right), \tag{4}$$

where V is as in (A).

Theorem 3.1. *Assume that P satisfies (A), $P^{(0)}$ satisfies (A0) and $|\omega|_{V^\alpha} < \infty$ for some $\alpha \in [0, 1/4)$. Then for any measurable function $f : \mathcal{X} \to \mathbb{R}$ such that $|f|_{V^\alpha} < \infty$,*

$$\mathbb{E}\left[f(X_n)|X_0 = x\right] \longrightarrow \pi(f), \quad as\ n \to \infty \tag{5}$$

and

$$\frac{1}{n}\sum_{i=0}^{n-1}\left(f(X_i) - \pi(f)\right) \xrightarrow{a.s.} 0, \quad as\ n \to \infty. \tag{6}$$

Proof. See Section 5. □

4 simulation examples

We illustrate the methods developed above with two examples from bayesian modelling. In the first example, we consider the Bayesian analysis of stochastic volatility models ((Kim et al., 1998)) and in the second example, we look at Bayesian phylogenetic trees reconstruction ((Larget and Simon, 1999)).

4.1 Bayesian analysis of stochastic volatility models

We consider the Bayesian analysis of the basic stochastic volatility model:

$$y_t = e^{h_t/2}\varepsilon_t, \quad t = 0, \ldots, T \tag{7}$$

$$h_{t+1} = \mu + \phi(h_t - \mu) + \sigma u_t, \quad t = 0, \ldots, T-1, \tag{8}$$

where (ε_t) and (u_t) are two uncorrelated sequences of i.i.d. standard normal random variables. We assume that $h_0 \sim N\left(\mu, \frac{\sigma^2}{1-\phi^2}\right)$ and $|\phi| < 1$ to assure the stationarity of

the process (h_t). We observe (y_t) but not (h_t), the so-called volatility process. The objective is to estimate $\theta = (\sigma, \phi, \beta)$ where $\beta = e^{\mu/2}$. This model and its generalizations have attracted attention in the financial econometrics literature as a better way to model financial markets series. A bayesian approach to analyze this model has been proposed by a number of authors (see e.g. (Kim et al., 1998) and the references therein). The difficulty is that the volatility process (h_t) is not observed making the likelihood of θ analytically intractable. The natural solution is to see (h_t) as a parameter and to design a Gibbs sampler on the posterior distribution $\pi(\theta, h_0, \ldots, h_T)$, of the parameter θ and the volatility process $(h_0, \ldots, h_T)$. But, due to the high autocorrelation in the volatility process, this sampler mixes very slowly. This mixing problem has motivated some authors to propose more sophisticated reparametrization of the model for better MCMC convergence. We show here that by resampling from the past in the Gibbs sampler, we can match the performances of the sophisticated solution proposed in (Kim et al., 1998).

We use the same prior distribution for θ as in (Kim et al., 1998) and essentially the same Gibbs sampler to sample from $\pi(\theta, h_0, \ldots, h_T)$ except when sampling from the conditional $\pi(h_t|\theta, h_{-t})$. To sample from this conditional, we use an Independent Metropolis sampler instead of the Accept-Reject method adopted in (Kim et al., 1998). The proposal distribution of our Independent Metropolis sampler is the same as the dominating distribution in the Accept-Reject sampler of (Kim et al., 1998). We refer the reader to (Kim et al., 1998) for the details.

Following (Kim et al., 1998) and (Shephard and Pitt, 1997), we use model (7) to analyze the Sterling dataset, which gives the daily observations of weekday close ex-

change rates for the UK Sterling/US Dollar exchange rate from 1/10/81 to 28/6/85. The total number of observations is $T = 946$. We first center the series with the formula

$y_t = 100\left[(\log(r_t) - \log(r_{t-1})) - \frac{1}{n}\sum_{j=1}^{n}(\log(r_j) - \log(r_{j-1}))\right]$, where (r_t) is the observed exchange rates. We then model (y_t) with the model (7).

We compare the plain Gibbs sampler with the 2 strategies discussed above: a Gibbs sampler with resampling from the past and a Gibbs sampler with resampling from an auxiliary process. To assure that the three sampler have about the same computational cost (storage requirement aside), we set the auxiliary process to be another copy of the plain Gibbs sampler with the same target distribution. The three samplers are run for $N = 250,000$ iterations. For each sampler and for each of the variables σ, ϕ, β, we give a plot of the last $5,000$ sample points together with the histogram and the autocorrelation function from the last $100,000$ points. When resampling from the past, the resampling schedule used is $B + \lceil k \rceil^{\alpha}$, $B = 125,000$ and $\alpha = 1.25$. For the third sampler with resampling from an auxiliary process, each of the two chains is run for $125,000$ iterations. The results of the variable σ (resp. ϕ and β) are given in in Graph 2 (resp. Graph 3 and Graph 4). On each graphics, the first column gives the result of the plain Gibbs sampler, the second column gives the results of the Gibbs sampler with resampling from the past and the results of the third sampler are in the third column.

Clearly, resampling from the past significantly improve on the Gibbs sampler. To quantify the gain, we compute, following (Kim et al., 1998) the inefficiency of each sampler on each of the three variables. For a Markov chain with transition kernel P

	σ	ϕ	β
Plain Gibbs	448.12	211.55	1.54
Gibbs with resampling	10.96	4.92	0.97
Gibbs with Aux. Proc.	12.24	9.91	1.39

Table 1: Inefficiencies of the samplers for the Sterling dataset.

and invariant distribution π, the inefficiency at f is:

$$I(f) = 1 + 2\sum_{k=1}^{\infty} \rho_k(f), \tag{9}$$

where $\rho_k(f) = Cov_\pi\left(f(X_k), f(X_0)\right) / Var_\pi\left(f(X_0)\right) = \pi\left(\bar{f} P^k \bar{f}\right) / \pi\left(\bar{f}^2\right)$. Basically, it is the cost of using a dependent process to sample from π. To estimate $I(f)$, we use, following Kim et al. (1998):

$$\hat{I}(f) = 1 + \frac{2B}{B-1} \sum_{i=1}^{B} K\left(\frac{i}{B}\right) \hat{\rho}_i(f), \tag{10}$$

where $\hat{\rho}_i(f)$ is the usual estimate of the autocorrelation at lag i for f and K the so-called Parzen kernel. We use $B = 5,000$. The result is given in Table 1.

By resampling from the past or from an auxiliary process, we obtain a sampler that outperforms (Shephard and Pitt, 1997) and is as efficient as the offset mixture method of (Kim et al., 1998).

4.2 Bayesian phylogeny reconstruction

Since Darwin's theory of evolution, methods to reconstruct the evolutionary relationships between different species have become important. We are concerned here

with the statistical inference of phylogenetic trees based on molecular sequences. Recently, more realistic models have been considered in this field owing to the MCMC machinery. We show here that MCMC samplers for phylogeny reconstruction can be improved upon with resampling from the past.

The statistical model is not standard, so we summarize it first. For more details on phylogenetic trees, we refer the reader to (Felsenstein, 2004). Suppose we have n aligned deoxyribonucleic acid (DNA) sequences $(y_1, \ldots, y_n)$ each of length m, where sequence i is from organism i. That is, $y_i = (y_i(1), \ldots, y_i(m))$ where $y_i(j)$ can be one of the four nucleotide basis A (Adenine), G (Guanine), C (Cytosine) or T (Thymine). Based on these sequences, we would like to infere the phylogenetic tree or evolutionary relationships between these organisms. To be precise, we recall that a binary tree τ for n species is a connected graph (V, E) with vertex set V and edges E, with no cycle, such that $V = \{\rho\} \cup \mathcal{I} \cup \mathcal{T}$, where ρ (the root) has degre 2; any $v \in \mathcal{I}$ has degre 3 and any $v \in \mathcal{T}$ has degre 1. $\mathcal{I}$ has $n-2$ elements called the internal nodes and $\mathcal{T}$ (the leaves or the tips) represent the n species. A phylogenetic tree for n species is a couple $\psi = (\tau, b)$, where τ is a binary tree for the n species and $b \in (0, \infty)^{|E|}$, where $|E| = 2n-1$ is the cardinality of E. For $e \in E$, b_e represents the length of edge e, the so-called branch length. We restrict our attention to phylogenetic trees with "contemporary tips", where the sum of the branch length b_e on the directed path from the root to any tip is constant (equal to 1 hereafter). Such phylogenetic trees are said to be with a "molecular clock" as the b_e can now be interpreted as time. Let Ψ be the set of all phylogenetic trees for n species. For $i \in V \setminus \{\rho\}$, denote $p(i)$ the parent of i, that is the vertex $p(i)$ such that $(p(i), i) \in E$.

The model of phylogenetic reconstruction we are interested in assumes that there are some missing DNA sequences $(y_j)_{\{j\in\{\rho\}\cup\mathcal{I}\}}$ such that the joint conditional distribution of $(y_j)_V$ given the phylogenetic tree ψ writes:

$$f((y_i)_{\{i\in V\}}|\psi) = f(y_\rho) \prod_{i\in V\setminus\{\rho\}} f\left(y_i|y_{p(i)},\psi\right). \tag{11}$$

In addition we make the simplifying assumption that each site evolves independently:

$$f(y_\rho) = \prod_{j=1}^{m} f(y_\rho(j)), \text{ and} \tag{12}$$

$$f\left(y_i|y_{p(i)},\psi\right) = \prod_{j=1}^{m} f\left(y_i(j)|y_{p(i)}(j), b_{(p(i),i)}\right). \tag{13}$$

And finally, we assume that there exist $(\pi_l)_{l\in\{A,G,C,T\}}$, $\pi_l \geq 0$, $\sum \pi_l = 1$, parameters $\theta, \kappa \in (0,\infty)$ and a 4×4 Markov process generator $Q = Q(\theta,\kappa,\pi_A,\pi_G,\pi_C,\pi_T)$ such that:

$$f(y_\rho(j) = l) = \pi_l,\ l\in\{A,G,C,T\} \text{ and} \tag{14}$$

$$f\left(y_i(j) = m|y_{p(i)}(j) = l, b_{(p(i),i)} = b\right) = exp(bQ)_{lm},\ \ l,m\in\{A,G,C,T\}. \tag{15}$$

The matrix Q specifies the model of DNA evolution. We use the F84 model as in (Larget and Simon, 1999). The parameters of the statistical model are then $(\psi,\theta,\kappa,\pi_A,\pi_G,\pi_C,\pi_T)$. To simplify the sampler, we fix π_A,π_G,π_C,π_T to their empirical values in the data. We assume that ψ has a uniform prior distribution on Ψ and we assume that θ and κ each has a uniform prior on $(0,M)$, $M = 200$. Let $\pi\left(\psi,\theta,\kappa|(y)_{i\in\mathcal{T}}\right)$ be the posterior distribution of the model. Clearly, $\pi\left(\psi,\theta,\kappa|(y)_{i\in\mathcal{T}}\right) \propto f\left((y)_{i\in\mathcal{T}}|\psi,\theta,\kappa\right)$ and this likelihood is obtained by integrating out the missing variables $(y_i)_{i\in\{\rho\}\cup\mathcal{I}}$ from (11). A fast computation of this likelihood is available with

the pruning method of Felsenstein (Felsenstein, 2004). To sample from this posterior distribution, we follow essentially (Larget and Simon, 1999). We update θ and κ together, given the phylogenetic tree ϕ, using a random walk Metropolis move. Next, given θ, κ, we update the phylogenetic tree ψ with the *global move with a molecular clock* of (Larget and Simon, 1999).

We compare this plain MCMC sampler with the samplers obtained with the two methods discussed in this paper. For the simulations, we use the *primate dataset* discussed in (Yang and Rannala, 1997). The dataset has $n = 9$ species and the phylogeny reconstruction is based on aligned sequences of length $m = 888$. The three samplers are simulated for $N = 500,000$ iterations. For each sampler and for each of the variables θ, κ, we give a plot of the last $5,000$ sample points together with the histogram and the autocorrelation function from the last $150,000$ iterations. When resampling from the past, the resampling schedule used is $B + \lceil k \rceil^{\alpha}$, $B = 100,000$ and $\alpha = 1.3$. For the third sampler with resampling from an auxiliary process, each of the two chains is run for $250,000$ iterations. The auxiliary process is a MCMC chain with stationary distribution $\pi^{(0)} = \pi^{1/T}$, with $T = 2$. The results of the variable θ (resp. κ) are given in in Graph 5 (resp. Graph 6). On each graphics, the first column gives the result of the plain MCMC sampler, the second column gives the results of the MCMC sampler with resampling from the past and the results of the third sampler are in the third column. In accordance with (Larget and Simon, 1999), the outputs of the three samplers overwhelmingly (with an estimated posterior distribution over 0.95) select the phylogenetic tree topology plotted in figure 7 as the most probable for this primate dataset.

	θ	κ
Plain MCMC	1510.23	1271.87
MCMC with resampling	13.95	24.37
MCMC with Aux. Proc.	9.18	8.15

Table 2: Inefficiencies of the samplers for the primates dataset

Here again, resampling from the past significantly improve on the plain MCMC sampler. Table 2 gives the efficiency gains.

5 Proofs of Theorem 2.1 and 3.1

We start with Theorem 2.1. Without any loss of generality we assume that B, the burn-in period is 0.

5.1 Proof of Theorem 2.1

The following lemma is a consequence of (A).

Lemma 5.1. *Assume (A). There exists a constant $C_1 \in (0, \infty)$ such that for any signed measure μ on $(\mathcal{X}, \mathcal{B})$ such that $\mu(\mathcal{X}) = 0$ and for any $n \geq 0$,*

$$\|\mu P^n\|_V \leq C_1 \rho^n \|\mu\|_V . \tag{16}$$

Proof of Theorem 2.1. Fix n such that $a_k \leq n < a_{k+1}$, $k \geq 2$. For $f \in L_V$ such that

$|f|_V \leq 1$, define $\bar{f} = f - \pi(f)$. We have:

$$\begin{aligned}
\mathbb{E}\left(\bar{f}(X_n)|X_0 = x\right) &= \mathbb{E}\left[\mathbb{E}\left(f(X_n)|X_{a_k}\right)|X_0 = x\right] \\
&= \mathbb{E}\left(P^{n-a_k}\bar{f}(X_{a_k})|X_0 = x\right) \qquad (17) \\
&= \left(\mathcal{L}^{(a_k)} - \pi\right)\left[P^{n-a_k}\bar{f}\,\right](x), \qquad (18)
\end{aligned}$$

where $\mathcal{L}^{(a_k)}(x, A) = \Pr\left(X_{a_k} \in A|X_0 = x\right)$. Therefore, since $\left|\!\left|\!\left|\mathcal{L}^{(n)} - \pi\right|\!\right|\!\right|_V = \sup_{x\in\mathcal{X}} \frac{\sup_{|f|_V\leq 1}\left|\mathbb{E}\left(\bar{f}(X_n)|X_0=x\right)\right|}{V(x)}$, it follows from Lemma 5.1, that:

$$\left|\!\left|\!\left|\mathcal{L}^{(n)} - \pi\right|\!\right|\!\right|_V \leq C_1\rho^{n-a_k}\left|\!\left|\!\left|\mathcal{L}^{(a_k)} - \pi\right|\!\right|\!\right|_V. \qquad (19)$$

Also, for $f \in L_V$ with $|f|_V \leq 1$, we have:

$$\begin{aligned}
\mathcal{L}^{(a_k)}\bar{f}(x) &= \mathbb{E}\left[\frac{1}{a_k}\sum_{j=0}^{a_k-1}\bar{f}(X_j)|X_0 = x\right] \\
&= \frac{a_{k-1}}{a_k}\mathcal{L}^{(a_{k-1})}\bar{f}(x) + \frac{1}{a_k}\sum_{j=a_{k-1}}^{a_k-1}\mathbb{E}\left(\bar{f}(X_j)|X_0 = x\right) \\
&= \frac{a_{k-1}}{a_k}\mathcal{L}^{(a_{k-1})}\bar{f}(x) + \frac{1}{a_k}\sum_{j=a_{k-1}}^{a_k-1}\left(\mathcal{L}^{(a_{k-1})} - \pi\right)P^{j-a_{k-1}}\bar{f}(x). \qquad (20)
\end{aligned}$$

Then proceding as above and using Lemma 5.1 again we get:

$$\left|\!\left|\!\left|\mathcal{L}^{(a_k)} - \pi\right|\!\right|\!\right|_V \leq \exp\left(-u_k\right)\left|\!\left|\!\left|\mathcal{L}^{(a_{k-1})} - \pi\right|\!\right|\!\right|_V, \qquad (21)$$

with $u_k = \log(a_k) - \log(a_{k-1} + c)$, $c = \frac{1}{1-\rho}$. If we define $u_1 = -a_1\log(\rho)$ and $\delta_k = \sum_{i=1}^{k} u_k$, we get $\left|\!\left|\!\left|\mathcal{L}^{(a_k)} - \pi\right|\!\right|\!\right|_V \leq C_2\exp(-\delta_k)$ for some finite constant C_2, which, together with (19) yields:

$$\left|\!\left|\!\left|\mathcal{L}^{(n)} - \pi\right|\!\right|\!\right|_V = O\left(\rho^{n-a_k}\exp(-\delta_k)\right), \qquad (22)$$

for $a_k \leq n < a_{k+1}$, as wanted.

□

5.2 Proof of Theorem 3.1

Let $\{X_n\}$ be the process generated by the importance-resampling scheme. We prove Theorem 3.1 as a consequence of Theorems 3.1 and 3.2 of (Atchade and Rosenthal, 2005). Denote $\mathcal{F}_n$ the σ-algebra generated by $(X_0, \ldots, X_n)$. For $x \in \mathcal{X}$ and $A \in \mathcal{B}$, define $P_n(x, A) = \Pr(X_n \in A | X_{n-1} = x) = \Pr(X_n \in A | \mathcal{F}_{n-1}, X_{n-1} = x)$. We have:

$$P_n(x, A) = \theta P(x, A) + (1 - \theta)\mu_n(A), \tag{23}$$

where $\mu_n(A) = \mathbb{E}\left[\frac{\sum_{k=0}^{n-1} \omega(X_k^{(0)}) \mathbf{1}_A(X_k^{(0)})}{\sum_{j=0}^{n-1} \omega(X_j^{(0)})}\right]$.

Define $M_r = \sup_n \mathbb{E}\left(V^r(X_n^{(0)})\right)$, $r \geq 0$. It follows from (A0) that $M_r \leq M_1 < \infty$ for all $r \in [0, 1]$. For $p \geq 0$, we write $\omega_i = \omega(X_i^{(0)})$, $s_n = \sum_{k=0}^{n-1} \omega_k$, $V_i^\alpha = V^\alpha(X_i^{(0)})$ and $\mu_n^{(p)} = \mathbb{E}\left[\frac{\sum_{i=0}^{n-1} \omega_i V_i^\alpha}{s_n}\right]^p$. The next lemma is crutial.

Lemma 5.2. *For $p \in [1, \frac{1}{4\alpha}]$, $\max_{0 \leq i \leq n-1} \mathbb{E}\left[\frac{\omega_i V_i^\alpha}{s_n}\right]^p = O\left(\frac{1}{n^p}\right)$ and $\mu_n^{(p)} = O(1)$ as $n \to \infty$.*

Proof. By the Minkowski inequality, we only need to prove that $\max_{0 \leq i \leq n-1} \mathbb{E}\left[\frac{\omega_i V_i^\alpha}{s_n}\right]^p = O\left(\frac{1}{n^p}\right)$.

Write $c = \lambda(h)$ and $c_0 = \lambda(h^{(0)})$. For $0 \leq i \leq n-1$ and $\kappa \in (0, c/c_0)$, we have:

$$\mathbb{E}\left[\frac{\omega_i V_i^\alpha}{s_n}\right]^p = \mathbb{E}\left[\frac{\omega_i V_i^\alpha}{s_n} \mathbf{1}_{\{s_n \geq n(c/c_0 - \kappa)\}}\right]^p + \mathbb{E}\left[\frac{\omega_i V_i^\alpha}{s_n} \mathbf{1}_{\{s_n < n(c/c_0 - \kappa)\}}\right]^p.$$

$$\mathbb{E}\left[\frac{\omega_i V_i^\alpha}{s_n} \mathbf{1}_{\{s_n \geq n(c/c_0 - \kappa)\}}\right]^p \leq \frac{1}{n^p (c/c_0 - \kappa)^p} \mathbb{E}\left[\omega_i V_i^\alpha\right]^p \leq \frac{|\omega|_{V^\alpha} M_1}{n^p (c/c_0 - \kappa)^p}.$$

By the Cauchy-Schwarz inequality, we can bound the second term as follows:

$$\begin{aligned}
\mathbb{E}\left[\frac{\omega_i V_i^\alpha}{s_n}\mathbf{1}_{\{s_n<n(c/c_0-\kappa)\}}\right]^p &\leq \mathbb{E}^{1/2}\left[\frac{\omega_i V_i^\alpha}{s_n}\right]^{2p}\left(\Pr\left[\frac{1}{n}\sum_{i=0}^{n-1}\left(\omega_i-\frac{c}{c_0}\right)<-\kappa\right]\right)^{1/2}\\
&\leq \mathbb{E}^{1/2}\left[V_i^{2p\alpha}\right]\left(\Pr\left[\frac{1}{n}\sum_{i=0}^{n-1}\left(\omega_i-\frac{c}{c_0}\right)<-\kappa\right]\right)^{1/2}\\
&\leq \frac{M_1^{1/2}}{n^{2p}\kappa^{4p}}\mathbb{E}^{1/2}\left(\sum_{i=0}^{n-1}\left(\omega_i-\frac{c}{c_0}\right)\right)^{4p},
\end{aligned}$$

where for the last line, the Markov inequality was used. Now we use the classical Poisson equation and martingale approximation technique. Since $\omega \leq V^\alpha$, the Poisson equation $\omega - c/c_0 = g - P^{(0)}g$ has a solution g which satisfies $|g| \leq V^\alpha$. With this solution, for $n > 1$, we can rewrite $\sum_{i=0}^{n-1}\omega_i - c/c_0 = M_n + W_n$ where $W_n = g(X_0^{(0)}) - P^{(0)}g(X_{n-1}^{(0)})$, $M_n = \sum_{i=1}^{n-1} g(X_i^{(0)}) - P^{(0)}g(X_{i-1}^{(0)})$ and (M_n) is a martingale. Therefore with the Minkowski inequality, we get: $\mathbb{E}^{1/2}\left(\sum_{i=0}^{n-1}(\omega_i - c/c_0)\right)^{4p} \leq \left[\mathbb{E}^{1/4p}\left(M_n\right)^{4p} + \mathbb{E}^{1/4p}\left(W_n\right)^{4p}\right]^{2p}$. Since $|g|_{V^\alpha} < \infty$ and $4p\alpha \leq 1$, it follows from Assumption (A0) that $\sup_{i,j}\mathbb{E}\left(g(X_i^{(0)}) - P^{(0)}g(X_j^{(0)})\right)^{4p} < \infty$. Therefore $\left(\mathbb{E}^{1/4p}\left(W_n\right)^{4p}\right)$ is bounded. Using Burkholder's inequality (see e.g. (Hall and Heyde, 1980)), we have the bound:

$$\begin{aligned}
\mathbb{E}\left(M_n\right)^{4p} &\leq K_3\mathbb{E}\left(\sum_{i=1}^{n-1}\left(g(X_i^{(0)}) - P^{(0)}g(X_{i-1}^{(0)})\right)^2\right)^{2p}\\
&\leq K_3\left[\sum_{i=1}^{n-1}\mathbb{E}^{1/2p}\left(g(X_i^{(0)}) - P^{(0)}g(X_{i-1}^{(0)})\right)^{4p}\right]^{2p}\\
&\leq K_4 n^{2p},
\end{aligned}$$

for some finite constants K_3, K_4. This implies that $\mathbb{E}^{1/2}\left(\sum_{i=0}^{n-1}(\omega_i - c/c_0)\right)^{4p} = O(n^p)$ which finishes the proof.

□

Lemma 5.3. *For all $n \geq 1$, P_n has an invariant distribution π_n, and for all $k \geq 0$,*

$$\left|\!\left|\!\left| P_n^k - \pi_n \right|\!\right|\!\right|_{V^\alpha} \leq C\theta^k \rho^k, \tag{24}$$

where the constant $C \in (0, \infty)$ does not depend on n or k. Moreover

$$\pi_n(f) \longrightarrow \pi(f), \quad \text{as } n \to \infty, \tag{25}$$

for any measurable function f, with $|f|_{V^\alpha} < \infty$.

Proof. One can directly check that the invariant distribution of P_n is π_n where:

$$\pi_n(A) = (1-\theta)\mu_n \left(\sum_{i=0}^{\infty} \theta^i P^i(x, A) \right). \tag{26}$$

And by recurrence, we can check that for $k \geq 0$ and $g \in L_{V^\alpha}$:

$$P_n^k g - \pi_n(g) = \theta^k P^k \bar{g} - (1-\theta)\mu_n \left(\sum_{i=k}^{\infty} \theta^i P^i \bar{g} \right). \tag{27}$$

Therefore $\left|\!\left|\!\left| P_n^k - \pi_n \right|\!\right|\!\right|_{V^\alpha} \leq \theta^k \rho^k \left(1 + \frac{1-\theta}{1-\theta\rho} \sup_n \mu_n(V^\alpha)\right)$ and according to Lemma 5.2,

$\sup_n \mu_n(V^\alpha)$ is finite.

For $f \in L_{V^\alpha}$, we write $\zeta(f) = (1-\theta)\sum_{i=0}^{\infty} \theta^i P^i f \in L_{V^\alpha}$. We have $|\pi_n(f) - \pi(f)| = \left|\mu_n\left(\zeta(\bar{f})\right)\right|$, where $\bar{f} = f - \pi(f)$. Note that $\pi(\zeta(\bar{f})) = 0$. We recall:

$$\mu_n\left(\bar{f}\right) = \mathbb{E}\left[\frac{\sum_{k=0}^{n-1} \omega(X_k^{(0)}) \bar{f}(X_k^{(0)})}{\sum_{j=0}^{n-1} \omega(X_j^{(0)})} \right]. \tag{28}$$

From the strong law of large numbers for $\{X^{(0)}\}$, the expression under the expectation in (28) converges a.s. to 0 as $n \to \infty$. On the other hand, for $p \in (1, 1/4\alpha)$,

$$\mathbb{E}\left| \frac{\sum_{k=0}^{n-1} \omega(X_k^{(0)}) \bar{f}(X_k^{(0)})}{\sum_{j=0}^{n-1} \omega(X_j^{(0)})} \right|^p \leq \mu_n^{(p)}, \tag{29}$$

and $(\mu_n^{(p)})$ is a bounded sequence. Therefore the sequence $\left(\frac{\sum_{k=0}^{n-1}\omega(X_k^{(0)})\bar{f}(X_k^{(0)})}{\sum_{j=0}^{n-1}\omega(X_j^{(0)})}\right)$ is uniformly integrable and it follows that $\mu_n(\bar{f}) \to 0$ as $n \to \infty$. □

Lemma 5.4.

$$\|\!|P_n - P_{n-1}|\!\|_{V^\alpha} + \|\pi_n - \pi_{n-1}\|_{V^\alpha} = O\left(\frac{1}{n}\right). \tag{30}$$

Proof. For $n \geq 1$, we have: $\|\!|P_n - P_{n-1}|\!\|_{V^\alpha} + \|\pi_n - \pi_{n-1}\|_{V^\alpha} \leq 2(1-\theta)\mathbb{E}\left[\frac{\omega_{n-1}V_{n-1}^\alpha}{\sum_{k=0}^{n-1}\omega_k}\right]$ and the lemma follows from Lemma 5.2. □

Proof of Theorem 3.1. Follows from Lemmas 5.3 and 5.4 and Theorems 3.1, 3.2 of Atchade and Rosenthal (2005). □

References

Atchade, Y. F. and Liu, J. S. (2006). Discussion of the paper by kou, zhou and wong. *Annals of Statistics* **To appear**.

Atchade, Y. F. and Rosenthal, J. S. (2005). On adaptive markov chain monte carlo algorithm. *Bernoulli* **11** 815–828.

Chauveau, D. and Vandekerkhove, P. (2001). Improving convergence of the hastings-metropolis algorithm with an adaptive proposal. *Scandinavian Journal of Statistics* **29** 13–29.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit theory and its application*. Academic Press, New York.

Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with arch models. *Review of Economic Studies* **62**.

Kou, S., Zhou, Q. and Wong, W. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Annals of Statistics* **To appear**.

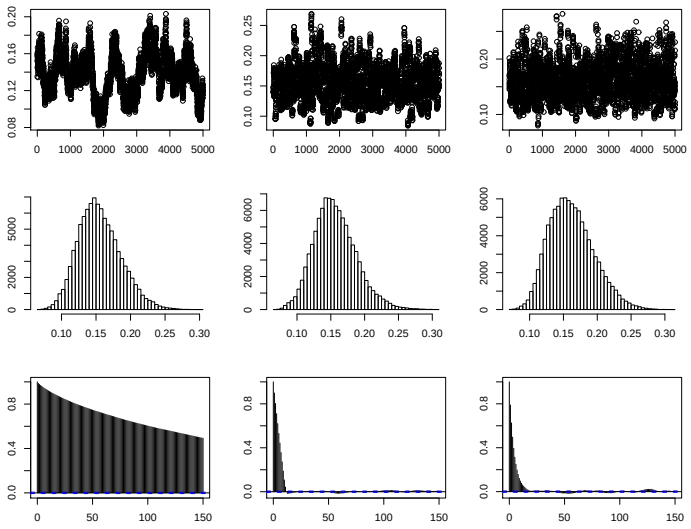
Larget, B. and Simon, D. L. (1999). Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16** 750–759.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer-Verlag London Ltd., London.

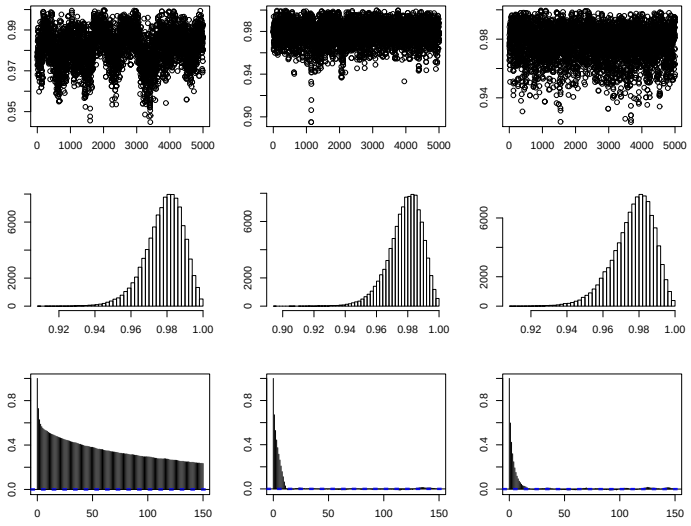
Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-gaussian measurement time-series. *Biometrika* **84** 653–667.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8** 1–9.

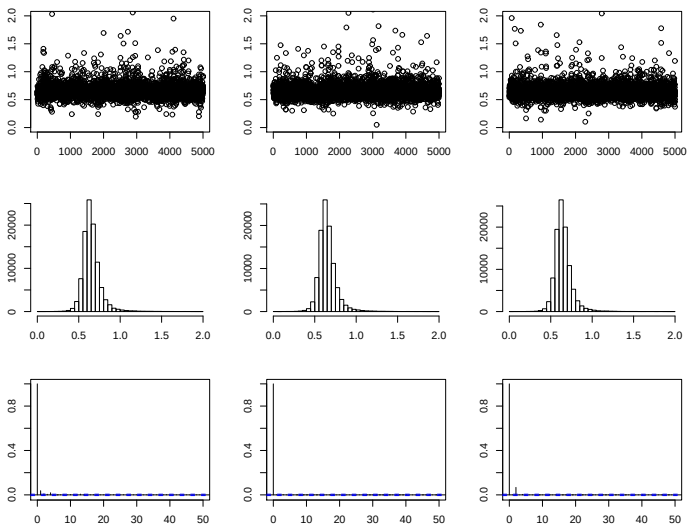
Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Mol. Biol. Evol.* **14** 717–724.



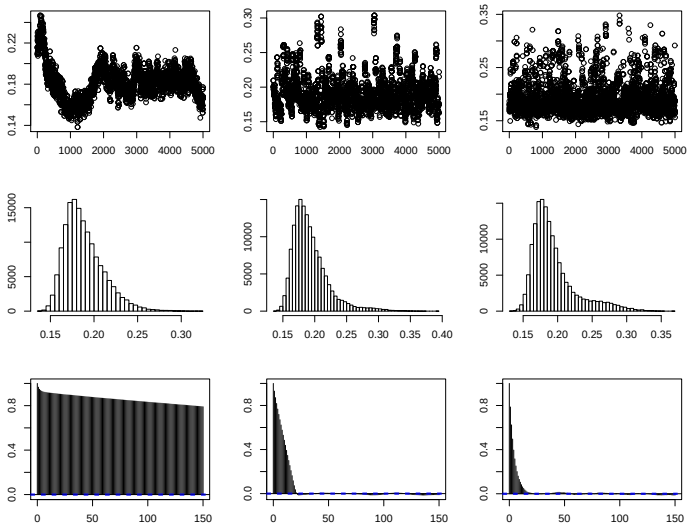
Graph 2: Outputs for σ. Sterling dataset. First column is the plain Gibbs, second column is resampling from the past; last column: resampling from an auxiliry Gibbs sampler.



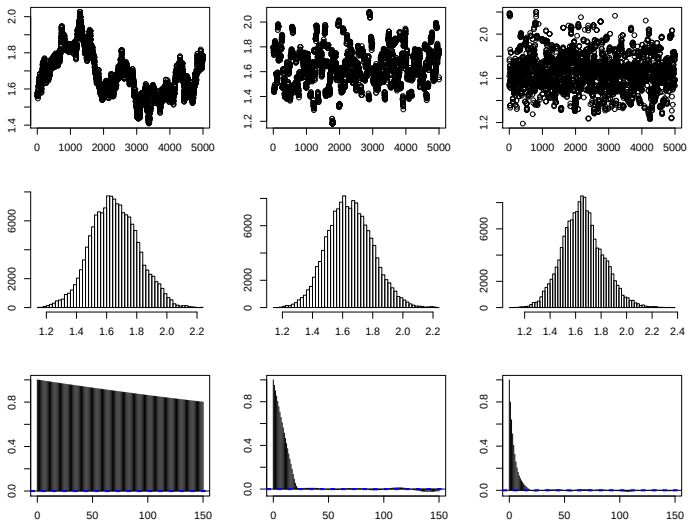
Graph 3: Outputs for ϕ. Sterling dataset. First column is the plain Gibbs, second column is resampling from the past; last column: resampling from an auxiliry Gibbs sampler.



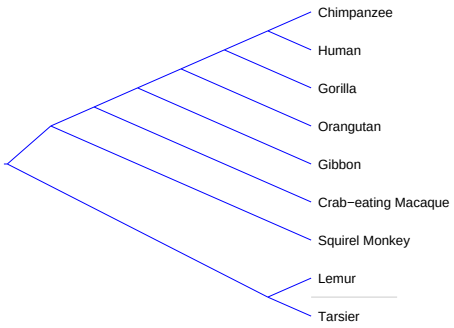
Graph 4: Outputs for β. Sterling dataset. First column is the plain Gibbs, second column is resampling from the past; last column: resampling from an auxiliary Gibbs sampler.



Graph 5: Outputs for θ. Primates dataset. First column is the plain MCMC, second column is resampling from the past; last column: resampling from an auxiliary MCMC sampler.



Graph 6: Outputs for κ. Primates dataset. First column is the plain MCMC, second column is resampling from the past; last column: resampling from an auxiliary MCMC sampler.



Graph 7: The most probable phylogenetic tree topology in the primates dataset.