Supplementary Document: Bayesian Phylogenetic Inference using a Combinatorial Sequential Monte Carlo Method

Liangliang Wang Alexandre Bouchard-Côté Arnaud Doucet

This Supplementary document assesses the correctness of our CSMC algorithm computer implementation using the joint distribution testing methodology of Geweke (2004). Since it is relatively easy to sample the evolutionary parameters in the particle MCMC algorithm, we assume that these parameters are known and omit them in the Geweke test. We will focus on testing the correctness of sampling phylogenetic trees.

In our Bayesian phylogenetic inference context, the Geweke test is based on two different ways to sample from the prior distribution $p(t)$ over tree t. In the Prior simulator, we sample t from $p(t)$. In the Prior-Posterior simulator, we first sample $\mathcal{Y}$ from $p(\mathcal{Y})$ by sampling t' from its prior distribution then $\mathcal{Y}$ from $\mathbb{P}(\mathcal{Y}|t')$, and we then sample t from the posterior $p(t|\mathcal{Y})$. This last simulation step is implemented using CSMC within particle MCMC (see Section 3.7 in the paper).

Hence if our computer implementation is correct then the samples t obtained from the Prior-Posterior simulator should be indistinguishable from draws from the Prior simulator. We compare statistically draws from the Prior and Prior-Posterior simulators using two test functions: the total tree length (sum of all branch lengths) and the diameter of the tree (the maximum number of edges between all pairs of leaves). The first test function tested the branch length distribution, and the second one tested the tree topology.

Ultrametric trees

As a sanity check, we voluntarily introduced the following temporary bugs in the proposal, prior, and the weight update to see if the Geweke tests failed in such cases:

1. A bug in proposing branch length: the tree rate was supposed to be 10, whereas it was changed to 1.

2. A bug in proposing topology: we assumed a uniform distribution over all pairs of subtrees in a forest, whereas we only considered the pairs of successive subtrees. That is, for a forest $s = \{(t_i, X_i)\}, i = 1, \cdots, |s|$, we randomly chose a pair to merge from pairs $\{(t_1, t_2), (t_2, t_3), \cdots, (t_{|s|}, t_1)\}$.

3. A bug in proposing both branch length and tree topology: a combination of Error 1 and Error 2.

4. A bug in the prior for branch lengths: the prior was supposed to be a coalescent tree (tree rate 10), whereas we used the exponential distribution with rate 10 for branch lengths.

5. A bug in the prior for tree topology: the prior was supposed to be a coalescent tree where we randomly chose a pair of subtrees in a forest to merge, whereas we changed this prior in a way to favour balanced trees. More specifically, for a forest $s = \{(t_i, X_i)\}, i = 1, \cdots, |s|$, the prior probability for a pair (t_i, t_j) was proportional to $|X_i| \cdot |X_j|$, where $|X_i|$ is the number of leaves in the subtree t_i.

6. A bug in the weight update: we changed the symbol of subtraction to addition when computing the log likelihood ratio.

	Tests (of 10 runs) for the total tree length failing at $p =$				Tests (10 runs) for the tree diameter failing at $p =$			
Error	.05	.01	.005	.001	.05	.01	.005	.001
None	0	0	0	0	0	0	0	0
1	10	10	10	10	0	0	0	0
2	3	1	0	0	10	10	10	10
3	10	10	10	10	10	10	10	10
4	10	10	10	10	6	5	5	2
5	1	1	1	1	5	2	1	0
6	6	3	2	1	2	0	0	0

Table 1: Summary of p-values of test statistics for the case of ultrametric trees.

Table 1 reports the number of rejections in the 10 independent runs of the Geweke test for the two test functions in the case of ultrametric trees, using some alternative conventional critical values. Both the Prior simulator and the Prior-Posterior simulator employed 1000 iterations. We used 10,000 particles in the CSMC algorithm used to propose trees for the particle MCMC scheme. The correct algorithm passed the joint distribution tests, whereas the 6 types of purposely introduced errors were all detected by at least one of the two tests.

Non-clock trees

1. A bug in proposing branch length: the rate in the exponential distribution was supposed to be 10, whereas we changed to $10 \cdot \binom{|s|}{2}$ where $|s|$ is the forest size in consideration.

2. A bug in proposing topology: the same type of bug as for ultrametric trees.

3. A bug in both proposing branch length and topology: a combination of Error 1 and Error 2.

4. A bug in the prior for branch lengths: the prior was supposed to be an exponential distribution of rate 10 for each branch, whereas we used the exponential

	Tests (of 10 runs) for the total tree length failing at $p =$				Tests (10 runs) for the tree diameter failing at $p =$			
Error	.05	.01	.005	.001	.05	.01	.005	.001
None	0	0	0	0	1	0	0	0
1	10	10	10	10	1	0	0	0
2	1	1	1	0	10	10	10	10
3	10	10	10	10	10	10	10	10
4	10	10	10	10	4	3	2	1
5	0	0	0	0	4	1	1	1
6	1	1	0	0	7	6	4	4

Table 2: Summary of p-values of test statistics for the case of non-clock trees.

distribution with rate 2.

5. A bug in the prior for tree topology: we were supposed to randomly choose a pair of subtrees in a forest to merge, whereas we changed this prior in a way to favour balanced trees. More specifically, for a forest $s = \{(t_i, X_i)\}, i = 1, \cdots, |s|$, the prior probability for a pair (t_i, t_j) was proportional to $|X_i| \cdot |X_j|$, where $|X_i|$ is the number of leaves in the subtree t_i.

6. A bug in the weight update: we omitted the backward proposal.

Table 2 shows the number of rejections in the 10 independent runs of the Geweke test for the two test functions in the case of non-clock trees, using some alternative conventional critical values. Both the Prior simulator and the Prior-Posterior simulator employed 1000 iterations. The correct algorithm passed the joint distribution tests, whereas the 6 types of purposely introduced errors were all detected by at least one of the two tests.

References

Geweke, J. (2004). Getting it right. *Journal of the American Statistical Association 99*(467), 799–804.