

Improved sampling scheme for BARCE

Dirk Husmeier

5 May 2004

The sampling scheme proposed in [2] is of the type of a Gibbs-within-Gibbs procedure, where the individual hidden states of the Markov chain, $\mathbf{S} = \{S_1, \dots, S_N\}$, are individually sampled in separate Gibbs steps by application of equation (17) in [2], which is itself a step within a Gibbs sampling procedure; see equation (13) of [2]. This procedure, which was first proposed in [4], shows slower mixing than sampling the whole state sequence \mathbf{S} from the posterior distribution directly. Let $P(\mathbf{S}|\mathcal{D})$ denote the probability distribution of state sequences conditional on the observed sequence alignment $\mathcal{D} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$, where \mathbf{y}_t denotes the column vector of nucleotides at site t in the alignment (the ‘observable’); the dependence of the distribution on the model parameters will not be made explicit in order to simplify the notation. Define

$$\alpha_t(S_t) = P(\mathbf{y}_1, \dots, \mathbf{y}_t, S_t) \quad (1)$$

which is the function computed in the forward algorithm of the forward-backward algorithm for hidden Markov models (HMMs); see, for instance, [3]. Now,

$$\begin{aligned} & P(S_t|S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) \\ \propto & P(S_t, S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) \\ = & P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+1}, \dots, S_N|S_t, \mathbf{y}_1, \dots, \mathbf{y}_t)P(S_t, \mathbf{y}_1, \dots, \mathbf{y}_t) \\ = & P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+1}, \dots, S_N|S_t)\alpha_t(S_t) \\ = & P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+2}, \dots, S_N|S_{t+1})P(S_{t+1}|S_t)\alpha_t(S_t) \\ \propto & P(S_{t+1}|S_t)\alpha_t(S_t) \end{aligned} \quad (2)$$

The simplifications carried out here follow directly from the independence relations in HMMs. The last step follows from the fact that the first term in the second-last line is independent of S_t and therefore cancels out in the normalization:

$$P(S_t = k|S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) = \frac{P(S_{t+1}|S_t = k)\alpha_t(S_t = k)}{\sum_i P(S_{t+1}|S_t = i)\alpha_t(S_t = i)} \quad (3)$$

Obviously, any scaling constant also cancels out in the normalization; hence replacing $\alpha_t(S_t)$ by some scaled version for numerical stabilization of the forward algorithm will

not affect the result. The algorithm is initialized by drawing the final state, S_N , from the following distribution:

$$P(S_N = k | \mathbf{y}_1, \dots, \mathbf{y}_N) = \frac{\alpha_N(S_N = k)}{\sum_i \alpha_N(S_N = i)} \quad (4)$$

The overall algorithm can thus be summarized as follows:

- Run the (scaled) forward-backward algorithm.
- Sample S_N from (4).
- Sample the remaining states S_{N-1}, \dots, S_1 recursively from (3).

It is straightforward to introduce a simulated annealing scheme during burn-in to accelerate convergence; see Section 6.5 in [1].

References

- [1] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK, 1998.
- [2] D. Husmeier and G. McGuire. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution*, 20(3):315–337, 2003.
- [3] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [4] C. P. Robert, G. Celeux, and J. Diebolt. Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics & Probability Letters*, 16:77–83, 1993.