Augmented Ensemble MCMC sampling in Factorial Hidden Markov Models

Kaspar Märtens
University of Oxford
kaspar.martens@stats.ox.ac.uk

Michalis K. Titsias
Athens University of Economics and Business
mtitsias@aueb.gr

Christopher Yau
Alan Turing Institute
University of Birmingham
c.yau@bham.ac.uk

Abstract

Bayesian inference for factorial hidden Markov models is challenging due to the exponentially sized latent variable space. Standard Monte Carlo samplers can have difficulties effectively exploring the posterior landscape and are often restricted to exploration around localised regions that depend on initialisation. We introduce a general purpose ensemble Markov Chain Monte Carlo (MCMC) technique to improve on existing poorly mixing samplers. This is achieved by combining parallel tempering and an auxiliary variable scheme to exchange information between the chains in an efficient way. The latter exploits a genetic algorithm within an augmented Gibbs sampler. We compare our technique with various existing samplers in a simulation study as well as in a cancer genomics application, demonstrating the improvements obtained by our augmented ensemble approach.

1 Introduction

Hidden Markov models (HMMs) are widely and successfully used for modeling sequential data across a range of areas, including signal processing (Crouse et al., 1998), genetics and computational biology (Marchini & Howie, 2010; Yau, 2013). The HMM assumes that there is an underlying unobserved Markov chain with a finite number of states, which generates a sequence of observations $y_{1:T} := (y_1, \dots, y_T)$ via a parametric emission distribution. Inference over the latent sequence $x_{1:T}$ and the parameters can be carried out either from a likelihood (Rabiner & Juang, 1986) or Bayesian (Scott, 2002) perspective. In the latter, conditional sampling can be used where the parameters and latent sequences are updated iteratively conditional on the other being fixed. Latent sequences can be sampled using forward-filtering-backward-sampling (FF-BS) (Scott, 2002).

The Factorial HMM (FHMM) (Ghahramani et al., 1997) is an extended version of the standard HMM where instead of a single latent chain, there are K latent chains. That is, given observations $y_{1:T}$, our goal is to infer a $K \times T$ latent matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ whose columns evolve according to Markov transitions. Here we focus on the case where $\mathbf{X}$ is binary, in which case the element $x_{k,t}$ indicates whether latent feature k contributes to observation y_t. The joint distribution $p(y_{1:T}, \mathbf{X})$ is given by

$$p(y_{1:T}, \mathbf{X}) = \left(\prod_{t=1}^{T} p(y_t|\mathbf{x}_t) \right) \left(p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) \right).$$

The FF-BS is an exact sampling algorithm and in principle, could be applied to FHMMs. However, this becomes infeasible even for a moderate number of latent sequences K. This is due to the state space growing exponentially with K. As the full FF-BS has complexity $O(2^{2K}T)$, a computationally cheaper approach is needed, however this comes at the expense of sampling efficiency.

One option is to sample each row of $\mathbf{X}$ conditional on the rest, using the FF-BS. Then each of the updates has a state space of size 2 and the FF-BS steps are inexpensive. However, in this conditional scheme most of the sequences are fixed and thus it is difficult for the sampler to explore the space well. A more general version of this would update a small subset of chains jointly at a higher computational cost, which can still get trapped in local modes.

An alternative idea referred to as Hamming Ball sampling has been suggested by Titsias & Yau (2014; 2017), which adaptively truncates the space via an auxiliary

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

variable scheme. Unlike the conditional Gibbs updates, it does not restrict parts of $\mathbf{X}$ to be fixed during sampling. Even though it can be less prone to get stuck, for a moderate value of K it may still not explore the whole posterior space.

This problem can be alleviated by ensemble MCMC methods which combine ideas from simulated annealing (Kirkpatrick et al., 1983) and genetic algorithms (Holland, 1992). One such example is parallel tempering (Geyer, 1991). Instead of running a single chain targeting the posterior, one introduces an ensemble of chains and assigns a temperature to each chain so that every chain would be targeting a tempered version of the posterior. Tempered targets are less peaked and therefore higher temperature chains in the ensemble explore the space well and do not get stuck. The key question becomes how to efficiently exchange information between the chains.

In this paper, we propose a novel ensemble MCMC method which provides an auxiliary variable construction to exchange information between chains. This is a general MCMC method, but our main focus is on improving existing poorly mixing samplers for sequence-type data. Specifically we consider the application to Factorial Hidden Markov Models. We demonstrate the practical utility of our augmented ensemble scheme in a series of numerical experiments, covering a toy sampling problem as well as inference for FHMMs. The latter involves a simulation study as well as a challenging cancer genomics application.

2 Augmented ensemble MCMC

Monte Carlo-based Bayesian inference (Andrieu et al., 2003) for complex high-dimensional posterior distributions is a challenging problem as efficient samplers need to be able to move across irregular landscapes that may contain many local modes (Gilks & Roberts, 1996; Frellsen et al., 2016; Betancourt, 2017). Commonly adopted sampling approaches can explore the space very slowly or become confined to regions around local modes.

Ensemble MCMC (also known as population-based MCMC, or evolutionary Monte Carlo) methods (Jasra et al., 2007; Neal, 2011; Shestopaloff & Neal, 2014) can alleviate this problem. This is achieved by introducing an ensemble of MCMC chains and then exchanging information between the chains. Next, we review standard ensemble MCMC approaches and proposal mechanisms that are used to exchange information. Then, we introduce our augmented Gibbs sampler.

2.1 Standard ensemble sampling methods

Suppose our goal is to sample from a density π. Instead of sampling $\mathbf{x} \sim \pi(\cdot)$, ensemble MCMC introduces an extended product space $(\mathbf{x}_1, \ldots, \mathbf{x}_K)$ with a new target density π^* defined as $\pi^*(\mathbf{x}_1, \ldots, \mathbf{x}_K) = \prod_{k=1}^K \pi_k(\mathbf{x}_k)$, where $\pi_k = \pi$ for at least one index. Here we focus on parallel tempering, which introduces a temperature ladder $T_1 < \ldots < T_K$ and associates a temperature with each chain. Denoting the inverse temperature $\beta_k = 1/T_k$, we define the tempered targets $\pi_k(\mathbf{x}_k) := \pi(\mathbf{x}_k)^{\beta_k}$. The idea is that high temperature chains can readily explore the space since the density is flattened by the power transformation, whereas the chain containing the true target density with $T_1 = 1.0$ only samples locally and precisely from the target. Each chain is updated independently, with occasional information exchange between the chains so that more substantial movement in the higher temperature chains can filter down to the slower moving low temperature chains.

One approach to exchanging information is to propose swapping states ("swap" move) between chains of consecutive temperatures and then performing an accept/reject operation according to the Metropolis-Hastings ratio (Geyer, 1991; Earl & Deem, 2005). However, a global move like this is unlikely to be accepted in a high-dimensional sampling setting.

More elaborate approaches can create proposals using genetic algorithms (Liang & Wong, 2000), by proposing certain moves between chains which again requires accepting/rejecting based on the Metropolis-Hastings framework. One such proposal scheme is a one-point crossover move, illustrated as follows:

$$\begin{matrix}(x_1, \ldots, x_t, x_{t+1}, \ldots, x_T) \\ (y_1, \ldots, y_t, y_{t+1}, \ldots, y_T)\end{matrix} \Longrightarrow \begin{matrix}(y_1, \ldots, y_t, x_{t+1}, \ldots, x_T) \\ (x_1, \ldots, x_t, y_{t+1}, \ldots, y_T)\end{matrix}$$

where the crossover point t could for example be chosen uniformly $t \in \{1, \ldots, T\}$. This is most natural for sequential models where there is dependency between consecutive x_t and x_{t+1}. For high-dimensional sequences this is more appealing than a swap move due to being more local and thus leading to higher chance of acceptance. One can similarly construct a two-point crossover move. However, the accept/reject procedure can be inefficient and very sensitive to *both* the choice of the temperature ladder and algorithmic parameter tuning. Our work seeks to address the latter issue by using an auxiliary variable augmentation that produces a Gibbs sampling scheme.

2.2 Gibbs sampling using auxiliary variables

Now, consider the target in the product space π^*. Suppose that during MCMC we would like to exchange

information between a pair of chains $\pi_i(\mathbf{x}_i)$ and $\pi_j(\mathbf{x}_j)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are T-dimensional vectors that indicate the current states of these chains. Here we describe an auxiliary variable move, which uses the idea of a one-point crossovers and leads to a Gibbs update for a two-point crossover.

We introduce two auxiliary variables $\mathbf{u}$ and $\mathbf{v}$, that live in the same space as $\mathbf{x}_i$ and $\mathbf{x}_j$, drawn from an auxiliary distribution $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$. Without loss of generality we assume that this auxiliary distribution is uniform over all possible one-point crossovers between $\mathbf{x}_i$ and $\mathbf{x}_j$.

We also introduce the set $\text{CR}(\mathbf{x}, \mathbf{y})$ to denote all T crossovers between the vectors $\mathbf{x}$ and $\mathbf{y}$. The auxiliary distribution $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ is precisely a uniform distribution over all pairs $(\mathbf{u}, \mathbf{v}) \in \text{CR}(\mathbf{x}_i, \mathbf{x}_j)$. This distribution is also symmetric, i.e. $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j) = p(\mathbf{x}_i, \mathbf{x}_j|\mathbf{u}, \mathbf{v})$.

Using the auxiliary variables we can exchange information between $\mathbf{x}_i$ and $\mathbf{x}_j$ through the intermediate step of sampling the auxiliary variables $(\mathbf{u}, \mathbf{v})$ based on the following two-step Gibbs procedure:

1. Generate $(\mathbf{u}, \mathbf{v}) \sim p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$

2. Generate $(\mathbf{x}_i, \mathbf{x}_j) \sim p(\mathbf{x}_i, \mathbf{x}_j|\text{rest})$, where

$$\begin{aligned} p(\mathbf{x}_i, \mathbf{x}_j|\text{rest}) &= \frac{1}{Z}\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j) \\ &= \frac{1}{Z}\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)p(\mathbf{x}_i, \mathbf{x}_j|\mathbf{u}, \mathbf{v}) \\ &= \frac{1}{Z}\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)I((\mathbf{x}_i, \mathbf{x}_j) \in \text{CR}(\mathbf{u}, \mathbf{v})) \end{aligned}$$

where the normalising constant $Z = Z(\mathbf{u}, \mathbf{v})$ is $Z(\mathbf{u}, \mathbf{v}) = \sum_{(\mathbf{y}_i, \mathbf{y}_j) \in \text{CR}(\mathbf{u}, \mathbf{v})} \pi_i(\mathbf{y}_i)\pi_j(\mathbf{y}_j)$.

The first step of the above procedure selects a random crossovered pair $(\mathbf{u}, \mathbf{v})$, while the second step conditions on this selected pair and jointly samples $(\mathbf{x}_i, \mathbf{x}_j)$ from the exact conditional posterior distribution that takes into account the information coming from the actual chains π_i and π_j.

Since the above is a Gibbs operation it leads to new state vectors for the chains π_i and π_j that are always accepted. To prove this explicitly we compute the effective marginal proposal and show that the corresponding Metropolis-Hastings acceptance probability is always one.

Given the current states $(\mathbf{x}_i, \mathbf{x}_j)$, we denote the proposed states by $(\mathbf{z}_i, \mathbf{z}_j)$ and the marginal proposal distribution by $Q(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)$. This proposal, defined by the above two-step Gibbs procedure, is a mixture:

$$\begin{aligned} &Q(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j) = \\ &= \iint \frac{1}{Z}\pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)p(\mathbf{z}_i, \mathbf{z}_j|\mathbf{u}, \mathbf{v})p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)d\mathbf{u}d\mathbf{v} \\ &= \pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)\iint \frac{1}{Z}p(\mathbf{z}_i, \mathbf{z}_j|\mathbf{u}, \mathbf{v})p(\mathbf{x}_i, \mathbf{x}_j|\mathbf{u}, \mathbf{v})d\mathbf{u}d\mathbf{v} \\ &= \pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)H(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j). \end{aligned}$$

The Metropolis-Hastings acceptance probability under this proposal is

$$\begin{aligned} \alpha &= \frac{\pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)Q(\mathbf{x}_i, \mathbf{x}_j|\mathbf{z}_i, \mathbf{z}_j)}{\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)Q(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)} \\ &= \frac{\pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)H(\mathbf{x}_i, \mathbf{x}_j|\mathbf{z}_i, \mathbf{z}_j)}{\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)\pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)H(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)}. \end{aligned}$$

Since $H(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j) = H(\mathbf{x}_i, \mathbf{x}_j|\mathbf{z}_i, \mathbf{z}_j)$ due to symmetry, all terms cancel out and $\alpha = 1$. So our proposal will be always accepted.

To simulate from $Q(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)$ in practice, we can use its mixture representation above, i.e. first generate auxiliary variables $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ and then conditional on those, generate the new value from $p(\mathbf{z}_i, \mathbf{z}_j|\mathbf{u}, \mathbf{v})$. We note that even though both of these steps are implemented as one-point crossovers, the overall proposal can lead to a two-point crossover as illustrated in Figure 1.

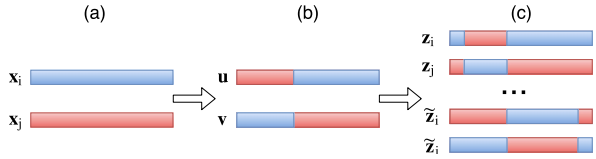

Figure 1: Schematic overview of the auxiliary variable crossover move. (a) We start with two sequences $\mathbf{x}_i$ and $\mathbf{x}_j$. (b) Now we construct auxiliary variables $\mathbf{u}, \mathbf{v}$ by applying a uniform one-point crossover to $\mathbf{x}_i, \mathbf{x}_j$. (c) Next, we consider all possible crossovers of $\mathbf{u}, \mathbf{v}$, and according to probabilities $\pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)$, we accept one of these configurations as the new value of $\mathbf{x}_i, \mathbf{x}_j$.

Specifically, to implement this, first we sample $(\mathbf{u}, \mathbf{v})$ uniformly from the set $\text{CR}(\mathbf{x}_i, \mathbf{x}_j)$. Now, conditional on the obtained $(\mathbf{u}, \mathbf{v})$, let us denote the crossover of $\mathbf{u}$ and $\mathbf{v}$ at point t by $(\mathbf{z}_i^{(t)}, \mathbf{z}_j^{(t)})$. The second step is to iterate over $t \in \{1, \ldots, T\}$ and compute quantities $a_t := \pi_i(\mathbf{z}_i^{(t)})\pi_j(\mathbf{z}_j^{(t)})$. The pair $(\mathbf{z}_i^{(t)}, \mathbf{z}_j^{(t)})$ will be accepted as the new value of $(\mathbf{x}_i, \mathbf{x}_j)$ with probability proportional to a_t.

A further extension of the above procedure is obtained by modifying the auxiliary distribution $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ to become uniform over the union of the sets $\text{CR}(\mathbf{x}_i, \mathbf{x}_j)$ and $\text{CR}(\mathbf{x}_j, \mathbf{x}_i)$ since, due to the deterministic ordering, the crossovers between $\mathbf{x}_i$ with $\mathbf{x}_j$ and the reverse crossovers between $\mathbf{x}_j$ with $\mathbf{x}_i$ are not identical.

The auxiliary distribution $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ still remains symmetric and all above properties hold unchanged. The only difference is that now we are considering $2T$ crossovers and in order to sample from $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ we need first to flip a coin to decide the order of $\mathbf{x}_i$ and $\mathbf{x}_j$. Complete pseudocode of the whole procedure is given in Supplementary.

The above sampling scheme is general and it can be applied to arbitrary MCMC inference problems involving both continuous and discrete variables. In the next section we apply the proposed method to a challenging inference problem in Factorial HMMs (FHMMs).

3 Application to FHMMs

Here, we apply the augmented ensemble scheme to FHMMs in order to improve on existing poorly mixing samplers. We achieve this via an ensemble of chains over suitably defined tempered posteriors. For a latent variable model, one can either temper the whole joint distribution or just the emission likelihood. We chose the latter, so the target posterior of interest becomes

$$\pi_k(\mathbf{X}) := p_k(\mathbf{X}|y_{1:T}) \propto p(\mathbf{X})p(y_{1:T}|\mathbf{X})^{\beta_k}$$

where $\mathbf{X}$ is a $K \times T$ binary matrix. As the ensemble crossover scheme was originally defined on vectors, there are multiple ways to extend this to matrices. One can perform crossovers on either rows or columns of a matrix, potentially considering a subset of those. Here we have decided to focus on a crossover move defined on the *rows* of $\mathbf{X}$, specifically on all K rows of $\mathbf{X}$.

The core computational step of the algorithm is to compute quantities a_t for all crossover points t. We show that these can be computed recursively in an efficient way. Let $\mathbf{U}$ and $\mathbf{V}$ be the current states of the auxiliary matrices for chains i and j. Comparing their crossovers at two consecutive points $t-1$ and t, denoted by $(\mathbf{Z}_{t-1}^{(i)}, \mathbf{Z}_{t-1}^{(j)})$ and $(\mathbf{Z}_{t}^{(i)}, \mathbf{Z}_{t}^{(j)})$, we note that these can differ just in column t:

$$\mathbf{Z}_{t-1}^{(i)} := (\mathbf{v}_1, \ldots, \mathbf{v}_{t-1}, \mathbf{u}_t, \mathbf{u}_{t+1}, \ldots, \mathbf{u}_T),$$
$$\mathbf{Z}_{t}^{(i)} := (\mathbf{v}_1, \ldots, \mathbf{v}_{t-1}, \mathbf{v}_t, \mathbf{u}_{t+1}, \ldots, \mathbf{u}_T).$$

As a result, the values $a_t = \pi_i(\mathbf{Z}_t^{(i)})\pi_j(\mathbf{Z}_t^{(j)})$ can be computed recursively. Indeed, given the previous value of $\pi_i(\mathbf{Z}_{t-1}^{(i)})$, we can compute $\pi_i(\mathbf{Z}_t^{(i)})$ by accounting for the following two cases: first, change in emission likelihood from $p(y_t|\mathbf{u}_t)^{\beta_i}$ to $p(y_t|\mathbf{v}_t)^{\beta_i}$, and second, change in the transitions from $\mathbf{v}_{t-1} \to \mathbf{u}_t \to \mathbf{u}_{t+1}$ to $\mathbf{v}_{t-1} \to \mathbf{v}_t \to \mathbf{u}_{t+1}$.

By denoting the overall transition probability $p(\mathbf{u}_{t+1}|\mathbf{u}_t)$ for chain i by $A^{(i)}(\mathbf{u}_t, \mathbf{u}_{t+1})$, we can express $\pi_i(\mathbf{Z}_t^{(i)})$ in terms of $\pi_i(\mathbf{Z}_{t-1}^{(i)})$ as follows $\pi_i(\mathbf{Z}_t^{(i)}) = \pi_i(\mathbf{Z}_{t-1}^{(i)}) \cdot c_t^{(i)}$ where

$$c_t^{(i)} := \frac{A^{(i)}(\mathbf{v}_{t-1}, \mathbf{v}_t)A^{(i)}(\mathbf{v}_t, \mathbf{u}_{t+1})}{A^{(i)}(\mathbf{v}_{t-1}, \mathbf{u}_t)A^{(i)}(\mathbf{u}_t, \mathbf{u}_{t+1})} \cdot \frac{p(y_t|\mathbf{v}_t)^{\beta_i}}{p(y_t|\mathbf{u}_t)^{\beta_i}}.$$

Now we can compute the quantities a_t recursively as follows $a_t = a_{t-1} \cdot c_t^{(i)} \cdot c_t^{(j)}$. As the values of a_t can be normalised to sum to one, we can arbitrarily fix the reference value $a_0 \leftarrow 1$. The computation of every correction term a_t is of the complexity $O(K)$, and the overall cost for all a_t values is $O(KT)$, being relatively cheap. As we typically need to perform the crossover moves only occasionally, the ensemble crossover scheme provides a way to improve the poorly mixing samplers for FHMMs at a small extra computational cost.

4 Experiments

First, we demonstrate the proposed sampling method on a multimodal toy inference problem. Then, we focus on Bayesian inference for FHMMs: we compare various samplers in a simulation study and then consider a challenging tumor deconvolution example. In both experiments, we compare a standard single-chain sampling technique (a Gibbs sampler or the Hamming Ball sampler) with the respective ensemble versions.

For ensemble samplers, we compare our proposed augmentation scheme ("augmented crossover") with two additional baseline exchange moves: the standard swap move ("swap") and a uniformly chosen crossover ("random cr") within the accept-reject Metropolis-Hastings framework. In all experiments, we run an ensemble of two MCMC chains, with temperatures $T_1 = 1.0$ and $T_2 = 5.0$, carrying out an exchange move every 10-th iteration.

4.1 Toy example

We consider the following multimodal toy sampling problem, where the target distribution is binary and has multiple separated modes. Specifically, we fix the dimensionality $T = 50$ and divide the sequence $\mathbf{x}$ into B contiguous blocks as follows $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(B)})$. In each of the blocks, we define a bimodal distribution, having two peaked modes $\mathbf{x}^{\text{mode}_1} := (1, 1, \ldots, 1)$ and $\mathbf{x}^{\text{mode}_2} := (0, 0, \ldots, 0)$, such that the probability of any binary vector $\mathbf{x}^{(j)}$ in block j is given by

$$p(\mathbf{x}^{(j)}) \propto \alpha_j^{\min(\mathrm{d}(\mathbf{x}^{(j)}, x^{\text{mode}_1}), \mathrm{d}(\mathbf{x}^{(j)}, \mathbf{x}^{\text{mode}_2}))} \quad (1)$$

where $\mathrm{d}(\cdot, \cdot)$ denotes the Hamming distance between two binary vectors and α_j is a block-specific parameter which controls how peaked the modes are. As a result, the further we go from the modes (in terms of Hamming distance), the less likely we are to observe that state. This has been illustrated in Figure 2.

We extend the above to define the joint $p(\mathbf{x})$ factorising

over the blocks as follows

$$p(\mathbf{x}) \propto \prod_{j=1}^{B} p(\mathbf{x}^{(j)}) = \prod_{j=1}^{B} \alpha_j^{\min(\mathrm{d}(\mathbf{x}^{(j)}, \mathbf{x}^{\mathrm{mode}_1}),\, \mathrm{d}(\mathbf{x}^{(j)}, \mathbf{x}^{\mathrm{mode}_2}))}$$

Within each block, the probability of a given state depends on its distance to the closest mode. This construction induces strong within-block dependencies. By varying the number of blocks within a sequence of fixed length, we can interpolate between a strong global correlation and local dependencies with a highly multimodal structure. The total number of modes for this distribution is 2^B, as illustrated in Figure 3.

In our experiments, we vary $B \in \{2, 5, 10\}$, resulting in distributions having $2^2, 2^5, 2^{10}$ modes. We generate $\alpha_j \in U(\{0.01, 0.02, \ldots, 0.05\})$. All samplers are initialised from the same value (one of the modes) and run for 10,000 iterations.

The resulting traces of $\mathbf{x}$ have been shown as heatmaps in Figure 4 for $B = 10$ (see Supplementary Figures for $B = 2$ and $B = 5$). As a summary statistic, we have shown the cumulative number of jumps between modes over repeated experiments in Figure 5 .

In all scenarios, the single chain Gibbs sampler expectedly struggles to escape the mode from which it was initialised, with ensemble methods better at moving between modes. For strong global correlations (corresponding to small B values), the baseline exchange moves "swap" and "random crossover" are reasonably efficient, though still result in a smaller number of mode jumps than the "augmented crossover".

Now when increasing B, the dependency structure becomes more local, resulting in a much more multimodal sampling landscape. For $B = 10$, the simple "swap" and "random crossover" moves struggle to accept any proposals at all and the benefit of our augmentation scheme becomes clear. In this highly multimodal setting with 2^{10} modes, the total number of modes visited by our "augmented crossover" (average 144) is much higher than for the "swap" (3) and "random crossover" (27) moves.

4.2 Tumor deconvolution example

The following example is motivated by an application in cancer genomics. Certain mutations in the cancer genome result in a loss of DNA integrity leading to copy number alterations due to the duplication or loss of certain DNA regions. Tumor samples consist of heterogeneous cell subpopulations and it is of interest to identify the subpopulations to study their phylogeny and gain insight into the clonal evolution (Ha et al., 2014; Gao et al., 2016). However, as DNA sequencing of bulk tissue samples produces aggregate data over all constituent cell subpopulations, the observed sequencing read counts must be deconvolved to reveal the underlying latent genetic architecture.

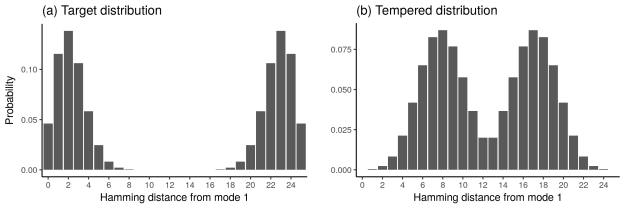

Figure 2: The bimodal within-block probability distribution as defined in eq. (1) for a binary sequence of length 25 shown in (a) and its tempered version in (b).

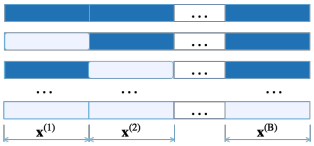

Figure 3: Multiple modes of the distribution of $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(B)})$, colour coding: dark = 1, light = 0.

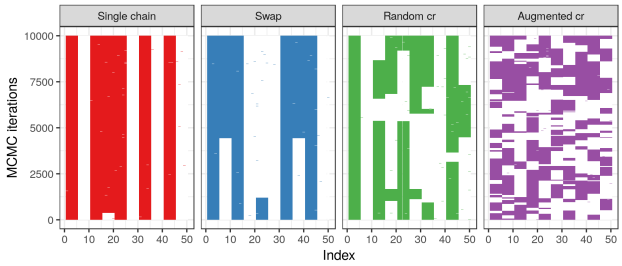

Figure 4: Heatmaps representing the trace plots of $\mathbf{x}$ for the experiment with $B = 10$ blocks, running a single chain Gibbs sampler (first panel), and its ensemble versions with various exchange moves: swap, random crossover, augmented crossover (in four panels). For each MCMC iteration, the elements of $\mathbf{x}$ have been colour coded: dark = 1, light = 0.

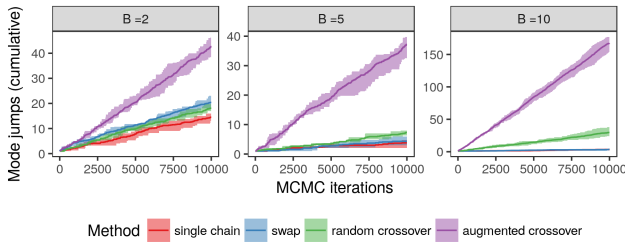

Figure 5: The cumulative number of jumps between modes (y-axis) over MCMC iterations (x-axis) for various experiments (block sizes $B \in \{2, 5, 10\}$) on average (coloured lines) together with 25% and 75% quantiles (shaded areas) over 10 repeated runs.

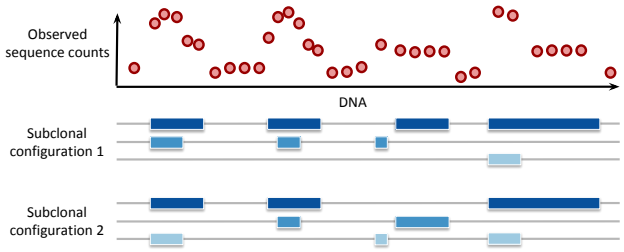

Figure 6: Existence of multiple subclonal configurations, both consisting of $K = 3$ subpopulations, which are indistinguishable when sequence counts (in top panel) are observed as aggregate over subpopulations.

The additive Factorial HMM is a natural model to consider where each latent chain corresponds to a putative cell subpopulation. However, it is important that the exploration of the state space of the latent chains allows us to identify the different subpopulation configurations that are compatible with the observed sequencing data since there maybe a number of plausible possibilities. This is illustrated in Figure 6. A poorly mixing sampler which is exploring only one of the possible latent explanations could lead to misleading conclusions regarding the subclonal architecture of a tumor. We wanted to examine if the ensemble scheme we propose could provide a more effective means of posterior sampling.

4.2.1 Simulation study

Lets consider the emission model $y_t|\mathbf{x}_t, \mathbf{w}, h \sim \mathcal{N}\left(h\sum_{k=1}^{K} w_k x_{k,t},\ \sigma^2\right)$ where y_t denote the observed sequence read counts at a locus t and h is the expected sequencing depth. Each w_k corresponds to the fraction of k-th subpopulation ($w_k \geq 0$, $\sum_k w_k = 1$) whose mutation profile is given by the k-th row of $\mathbf{X}$. Here $x_{k,t} \in \{0, 1\}$ denotes whether the k-th population has a copy number alteration at position t or not.

Note that this is not a complete model of real-world sequencing data but a simplified version to demonstrate the utility of the proposed ensemble MCMC methods. The results presented here should extend to the more complex cases. Further work to construct a sufficiently complex model to capture the variations within real sequencing data, such as single nucleotide polymorphisms, is beyond the scope of this paper and will be developed in future work.

First, we investigated the performance of sampling schemes for FHMMs in the presence of multimodality in a controlled setting. We generated observations from the emission distribution with $K = 3$ with weights such that $w_1 + w_2 \approx w_3$. As a result, data generating scenarios $\mathbf{x}_t = (1, 1, 0)$ and $\mathbf{x}_t = (0, 0, 1)$ are both plausible underlying latent explanations.

For data generation, we used a latent $\mathbf{X}$ matrix having a block structure of columns $(1, 1, 0)$ followed by a block of $(0, 0, 0)$, as illustrated in Figure 7(a), but using altogether 20 blocks. We fixed $h = 15$, $\mathbf{w} = (0.2+\varepsilon, 0.3+\varepsilon, 0.5-2\varepsilon)$ with $\varepsilon = 0.01$ and $\sigma^2 = 1$. Each of these blocks has two modes, but due to the structured FHMM prior on $\mathbf{X}$, the mode $(0, 0, 1)$ corresponds to a slightly higher log-posterior value. For example, the three examples provided in Figure 7 are ordered in terms of posterior probability (c) > (b) > (a).

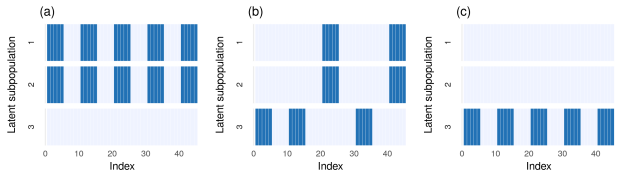

Figure 7: Small illustration of three possible modes for the $\mathbf{X}$ matrix used in the simulation study.

For inference in FHMMs, we considered two single chain samplers for $\mathbf{X}$: one-row updates conditional on the rest ("Gibbs"), and the Hamming Ball sampler ("HB"). We then considered ensemble versions of both of these samplers, as shown in Figure 8 (left column for "Gibbs" and right column for "HB"). All chains were initialised from the mode with $x_{3,t} = 0$, i.e. mode (a) in Figure 7, and were ran for 10 000 iterations. Exchange moves were carried out every 10th iteration.

For "Gibbs", the single chain sampler and the "swap" ensemble have not moved from the initialisation, the "random cr" ensemble scheme shows some improvement, but the "augmented cr" has quickly moved towards values of $\mathbf{X}$ with higher posterior probability (see Figure 8(a)). It also exhibits much better mixing, as seen from the traces of the first row of $\mathbf{X}$, i.e. traces of $x_{1,1:T}$ shown in Figure 8(c). We note that $x_{1,t} = 0$ values correspond to the more probable mode.

As a single chain sampler, "HB" quickly achieves higher log-posterior values than "Gibbs". Therefore, for "HB" the gain from "swap" and "random cr" ensemble techniques is relatively smaller, but still the "augmented cr" has quickly moved towards higher log-posterior values.

To quantify mixing on binary state spaces, we have calculated the Hamming distance between $\mathbf{X}^{(t)}$ and $\mathbf{X}^{(t+\text{lag})}$ for various lag values $\{1, 10, 50\}$, normalised by $\dim(\mathbf{X})$. Panels (e, f) show the distribution of these summary statistics, confirming that the augmented crossover scheme reduces notably the dependence between consecutive samples of $\mathbf{X}$.

We have shown above that the complexity of augmented

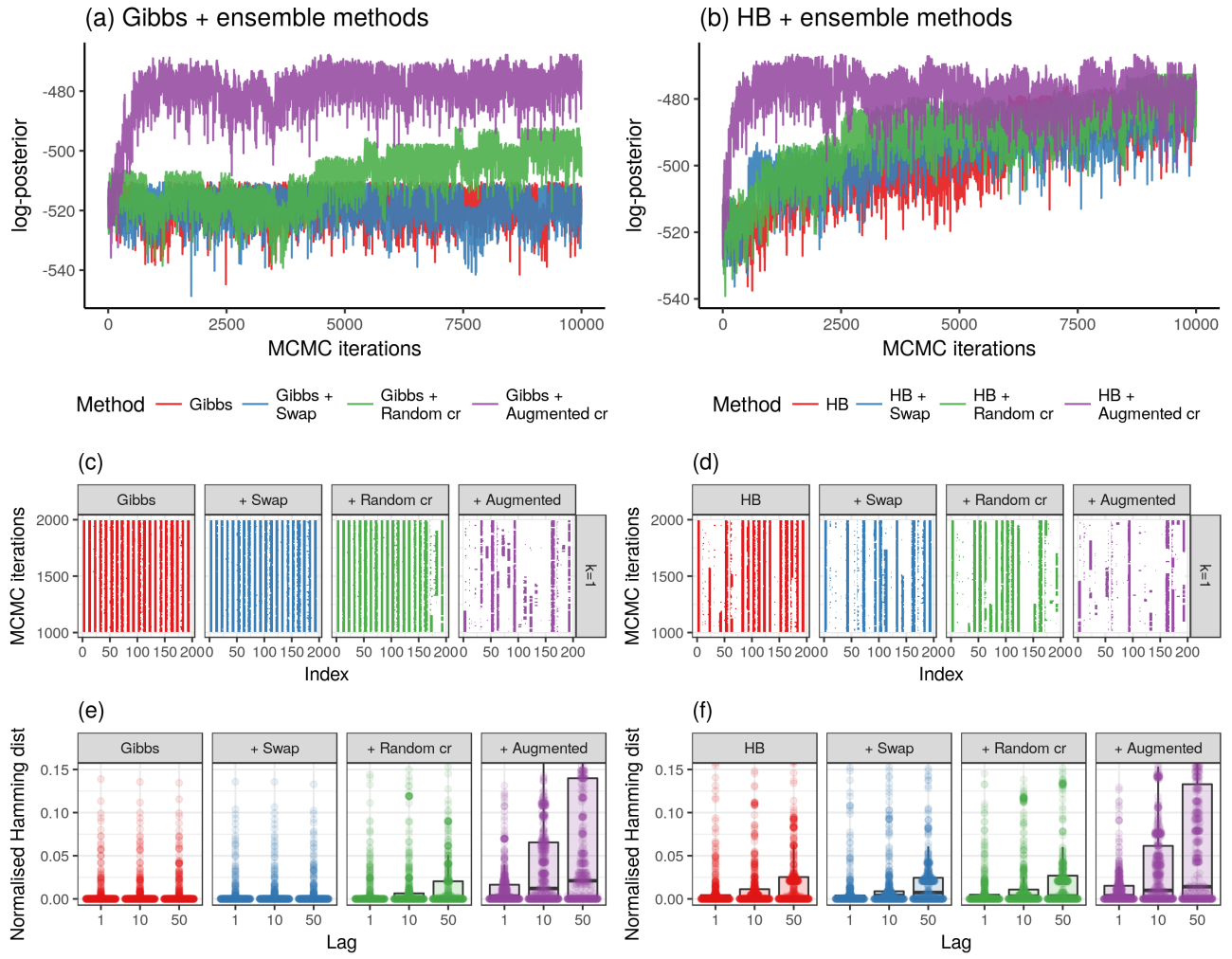

Figure 8: Simulation study comparing sampling techniques for FHMMs: ensemble versions of the Gibbs sampler (left column) and of the Hamming Ball sampler (right column). (a, b) Traces of log-posterior for the single chain sampler ("Gibbs", "HB") and three ensemble versions. (c, d) Heatmaps showing the traces for the first row ($k = 1$) of $\mathbf{X}$ (colour coded: dark = 1, light = 0), zoomed in to MCMC iterations 1000 - 2000 (y-axis). (e,f) Distribution of the normalised Hamming distance: boxplots in the background, overlaid with individual values) for various lags 1, 10, 50 (x-axis).

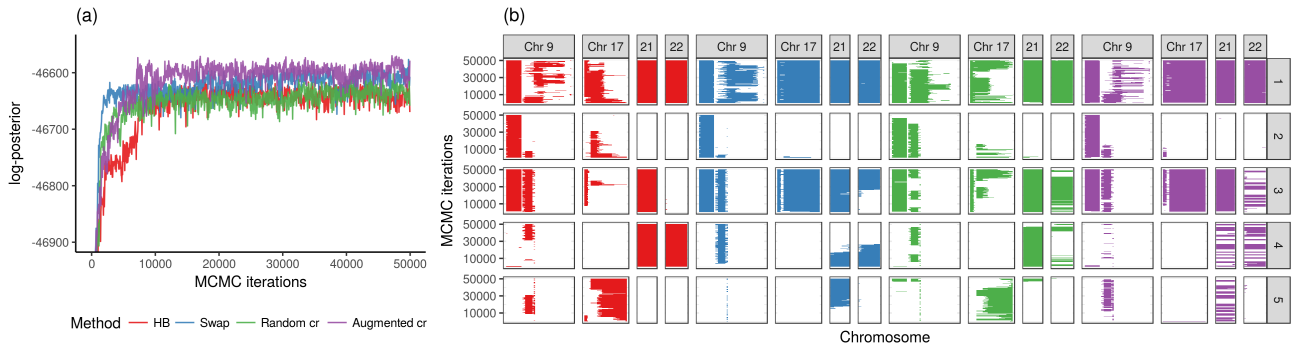

Figure 9: FHMM fitted to real sequencing data, using HB($r = 3$) single-chain sampler and the corresponding ensemble samplers ("swap", "random cr", "augmented cr"). (a) Traces of log-posterior (y-axis), and (b) traces of $\mathbf{X}$ over MCMC iterations for each 5 rows of $\mathbf{X}$ (row panels 1 – 5), with the genomic coordinates (x-axis) zoomed in to selected chromosomes.

Table 1: Computation times in seconds for the simulation study (two chains, 10 000 iterations).

HB	swap	random cr	augmented cr
130	132	133	135
Gibbs	swap	random cr	augmented cr
213	216	214	218

crossover scheme is linear $O(KT)$, which is also the case for the "swap" and "random cr" moves. To explore the respective costs in practice, we measured the total computation time for our Rcpp implementation. To establish the baseline cost of running a two-chain ensemble without any exchange moves in a sequential implementation, we indicate this baseline time in the first column ("Gibbs" and "HB") of Table 1. We note that this could be halved by a parallel implementation. The extra cost for all exchange moves are relatively small. Even though the extra time for the "swap" and "random cr" schemes is just slightly smaller than for "augmented cr", this is a small price to pay for an improvement in mixing, especially compared to the high baseline cost of running an FHMM sampler.

4.2.2 Tumor data analysis

Next we consider whole-genome tumor sequencing data for bladder cancer (Cazier et al., 2014). To illustrate the utility of our sampling approach, we used data from one patient (patient ID: 451) and took a thinned sample of 10,877 loci. We placed a vague Gaussian prior on the expected sequencing depth, $h \sim \mathcal{N}(\mu_h, \sigma_h^2)$ with $\mu_h = 180, \sigma_h = 30$, and integrated out h, resulting in the marginal likelihood

$$y_t|\mathbf{x}_t, \mathbf{w} \sim \mathcal{N}\left(\mu_h \sum_{k=1}^{K} w_k x_{k,t},\ \sigma^2 + \sigma_h^2 \left(\sum_{k=1}^{K} w_k x_{k,t}\right)^2\right).$$

Here each row of $\mathbf{X}$ corresponds to a single chromosome and the binary state indicates whether a copy of that DNA region exists or not. We fixed $K = 6$, where one of the latent sequences is always fixed to 1, representing a baseline, unaltered chromosome. We used a Hamming Ball Sampler with radius $r = 3$ as a single chain sampler, and its tempered ensemble versions "swap", "random cr", and "augmented cr".

Since it is the sampling efficiency of the latent chains $\mathbf{X}$ in the FHMM rather than associated parameters that is the direct target of our sampler, we fixed $\mathbf{w}$ value to $(0.075, 0.125, 0.15, 0.175, 0.2, 0.275)$ in these experiments. As a result, all samplers would be exploring the same conditional posterior, and we are able to directly compare the subclonal configurations identified by various sampling algorithms. Otherwise, joint updating of the weights $\mathbf{w}$ (though entirely feasible) would lead to label swapping effects and the possibility of samplers exploring entirely different regimes that then make direct comparisons across sampling methods more challenging.

Figure 9 shows the log-posterior traces and the traces of $\mathbf{X}$ for selected chromosomes, when using ensembles of the HB($r = 3$) sampler. After a burn-in period of 10 000 iterations, the "augmented cr" ensemble has identified a probable configuration of $\mathbf{X}$ and it continues to explore parts of the state space which have higher posterior probability than those identified by other samplers.

The augmented sampler is much better at capturing the uncertainty in underlying latent configurations (see Figure 9(b)). For example, the third row corresponds to a subpopulation which has an extra copy of chromosome 21, but there is uncertainty whether it co-occurs with a whole extra copy of chromosome 22. Examining chromosome 17, the single-chain HB sampler and the "random cr" ensemble have identified a more fragmented latent configuration, whereas "swap" and "augmented cr" have combined these fragments into an alternative, more probable explanation. In biological terms, this is important since the more fragmented configuration would suggest a highly genomically unstable cancer genome related to a loss of genome integrity checkpoint mechanisms, whilst the alternative suggests a more moderate degree of instability.

5 Conclusion

We introduce an ensemble MCMC method to improve poorly mixing samplers for FHMMs. This is achieved by combining parallel tempering and a novel exchange move between pairs of chains achieved through an auxiliary variable augmentation. The former introduces a chain which explores the space freely and does not get stuck, whereas the latter provides an efficient procedure to exchange information between a tempered chain and our target. The proposed method is a general purpose ensemble MCMC approach, but its most natural application case are sequential models. Specifically, we see this most useful for a broad class of models assuming Markov structure, where the augmented crossover move can be carried out at a cheap extra computational cost. A natural extension of this work is to integrate our ensemble technique into a sampling scheme for targeting latent variables $\mathbf{X}$ *and* parameters θ in a joint model $\pi(\mathbf{X}, \theta)$. More exploration could also be carried out to explore optimal strategies for selecting or adapting the temperature ladder. However, our analyses suggest that for any given temperature ladder, the suggested augmented crossovers outperform non-augmented, classic approaches.

Acknowledgements

KM is supported by a UK Engineering and Physical Sciences Research Council Doctoral Studentship. CY is supported by a UK Medical Research Council Research Grant (Ref: MR/P02646X/1) and by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Andrieu, Christophe, De Freitas, Nando, Doucet, Arnaud, and Jordan, Michael I. An introduction to MCMC for machine learning. *Machine learning*, 50 (1-2):5–43, 2003.

Betancourt, Michael. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Cazier, J-B, Rao, SR, McLean, CM, Walker, AK, Wright, BJ, Jaeger, EEM, Kartsonaki, C, Marsden, L, Yau, C, Camps, C, et al. Whole-genome sequencing of bladder cancers reveals somatic cdkn1a mutations and clinicopathological associations with mutation burden. *Nature communications*, 5:3756, 2014.

Crouse, Matthew S, Nowak, Robert D, and Baraniuk, Richard G. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on signal processing*, 46(4):886–902, 1998.

Earl, David J and Deem, Michael W. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Frellsen, Jes, Winther, Ole, Ghahramani, Zoubin, and Ferkinghoff-Borg, Jesper. Bayesian generalised ensemble Markov chain Monte Carlo. In *Artificial Intelligence and Statistics*, pp. 408–416, 2016.

Gao, Ruli, Davis, Alexander, McDonald, Thomas O, Sei, Emi, Shi, Xiuqing, Wang, Yong, Tsai, Pei-Ching, Casasent, Anna, Waters, Jill, Zhang, Hong, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature Genetics*, 2016.

Geyer, CJ. *Computing Science and Statistics Proceedings of the 23 Symposium on the Interface; American Statistical Association: New York; p 156*, 1991.

Ghahramani, Zoubin, Jordan, Michael I, and Smyth, Padhraic. Factorial hidden Markov models. *Machine learning*, 29(2-3):245–273, 1997.

Gilks, Walter R and Roberts, Gareth O. Strategies for improving MCMC. *Markov chain Monte Carlo in practice*, 6:89–114, 1996.

Ha, Gavin, Roth, Andrew, Khattra, Jaswinder, Ho, Julie, Yap, Damian, Prentice, Leah M, Melnyk, Nataliya, McPherson, Andrew, Bashashati, Ali, Laks, Emma, et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research*, 24 (11):1881–1893, 2014.

Holland, John H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

Jasra, Ajay, Stephens, David A, and Holmes, Christopher C. On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279, 2007.

Kirkpatrick, Scott, Gelatt, C Daniel, Vecchi, Mario P, et al. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

Liang, Faming and Wong, Wing Hung. Evolutionary Monte Carlo: Applications to cp model sampling and change point problem. *Statistica sinica*, pp. 317–342, 2000.

Marchini, Jonathan and Howie, Bryan. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

Neal, Radford M. MCMC using ensembles of states for problems with fast and slow variables such as gaussian process regression. *arXiv preprint arXiv:1101.0387*, 2011.

Rabiner, Lawrence and Juang, B. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1): 4–16, 1986.

Scott, Steven L. Bayesian methods for hidden Markov models. *Journal of the American Statistical Association*, 2002.

Shestopaloff, Alexander Y and Neal, Radford M. Efficient bayesian inference for stochastic volatility models with ensemble MCMC methods. *arXiv preprint arXiv:1412.3013*, 2014.

Titsias, Michalis K and Yau, Christopher. Hamming ball auxiliary sampling for factorial hidden Markov models. In *Advances in Neural Information Processing Systems*, pp. 2960–2968, 2014.

Titsias, Michalis K and Yau, Christopher. The hamming ball sampler. *Journal of the American Statistical Association*, pp. 1–14, 2017.

Yau, Christopher. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics*, 29(19):2482–2484, 2013.