Change Points in Finite-state, Piece-Wise Markov Chains

James Thornton

November 1, 2018

Abstract

This paper explores two complementary methods of identifying the change-points in a finite-state, piece-wise Markov Chain: one Bayesian, one with a frequentist interpretation. Each with advantages and disadvantages based on context. Empirical results on artificial data with known change-points will be briefly discussed.

1 Introduction

Genomic data can be viewed as a lengthy sequence of letters {"G", "T", "A", "C"}, known as nucleotides. In a simple abstraction, this sequence of letters can be modelled as a piece-wise concatenation of realised finite-state Markov Chains, each with possibly different transition matrices and unknown length. It is hoped that "pieces" of the concatenated chain corresponding to different regions of chromosome. This motivates the general problem of a sequential clustering or change-point analysis on piece-wise, finite-state Markov Chains.

There are a number of approaches one may take to perform this clustering: maximum likelihood estimation (MLE) on the assumed data generating process, a distance based clustering heuristic, or perhaps a latent state model on the clusters and associated transition matrices. Such latent state models may include a Hidden Markov Model (HMM) or non-parametric model such as using the Dirichlet Process to control for the unknown number of clusters. This paper shall however focus on a Bayesian model and then a heuristic with interesting probabilistic intuition.

2 Theory

2.1 Data Generating Process

Let $x_{1:N} = (x_1, x_2, \ldots, x_N)$ denote the observed, realised chain of states, corresponding to the random variables $X_{1:N}$, where $x_i \in [1:m]$ $\forall i$, and $[1:m]$ denotes set $\{1, 2, \ldots, m\}$. Given our assumed piece-wise Markov Chain structure, let K be the number of chain-pieces, and $M_k = \{m_{k,i,j} | i \in [1:m], j \in [1:m]\} \in R^{m \times m}$ be the underlying transition matrix that generated the k^{th} chain-piece, where $k \in [1:K]$. $\mathcal{M} = \{M_k | k \in [1:K]\}$. Encode the change points as $\mathcal{C} = \{c_1, c_2, \ldots, c_K\} \in [1:N]^K$ and denote indicators $\mathcal{S} = \{s_i \in [1:K] |\ i \in [1:N]\}$, where $s_i = k$, if the i^{th} entry of the chain is in the k^{th} cluster. Finally denote π to be the unconditional distribution on X_1, however this is of little consequence in our analysis. For brevity denote $\theta = (\mathcal{M}, \mathcal{S}, \mathcal{C}, K)$, and let the likelihood of the observed data be $\mathcal{L}$, whereby:

$$\mathcal{L}(\theta, x_{1:N}) = \pi(x_1) \prod_{k=1}^{K} \prod_{\substack{i=1 \\ s_i = k}}^{N-1} m_{k,x_i,x_{i+1}} \tag{1}$$

Alternatively, let $\mathcal{T} = \{t \in [1:m]^2\}$ be the set of transitions and let $F_t^k(x_{1:N}, \mathcal{S}) = F_t^k$ denote the frequency of transitions of type $t \in \mathcal{T}$ for the subset of the chain $x_{1:N}$ where $s_i = k$ for $i \in [1:N]$, then:

$$\mathcal{L}(\theta, x_{1:N}) = \pi(x_1) \prod_{k=1}^{K} \prod_{i=1}^{m} \prod_{j=1}^{m} m_{k,i,j}^{F_{i,j}^k} \tag{2}$$

Consider a naive, frequentist, MLE approach where there exists some 'true' parameters, θ, yet no prior information on the number of chain-pieces. One could choose the optimal change-points as the chain index itself, and for each of the N chain-pieces set the transition probability to be 1 at the observed transition. This is clearly not useful, suggesting some constraint or prior information is required over the number of change-points. A natural set of enhancements to the MLE approach, all of which may possibly have some deeper underlying equivalence, would be: a Bayesian approach with a prior on the number of change points, a penalised likelihood method or some constraint on the parameter space. Indeed the remainder of this report will primarily explore a Bayesian approach, and also a simple grid-merge heuristic which has some interesting probabilistic properties.

2.2 Bayesian Conjugate Model

Let $Dir(\boldsymbol{\alpha})$ denote the Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, and density $dir(\cdot, \boldsymbol{\alpha})$. Define distribution $DirMat(\boldsymbol{\alpha})$, where now $\boldsymbol{\alpha} \in R^{m \times m}$, as follows. If $M \sim DirMat(\boldsymbol{\alpha})$ then $M_i \overset{i.i.d}{\sim} Dir(\boldsymbol{\alpha}_i)$ where B_i is the i^{th} row of matrix B. $DirMat$ has density denoted $dirmat(\cdot, \boldsymbol{\alpha})$ where:

$$dirmat(R, \boldsymbol{\alpha}) \propto \prod_{i=1}^{m} dir(R_i, \boldsymbol{\alpha}_i) \propto \prod_{i=1}^{m} \prod_{j=1}^{m} r_{i,j}^{\alpha_{i,j}-1} \tag{3}$$

Consider the allocation parameter, $\mathcal{S}$. Without loss of generality, one can set $s_{i+1} \in \{s_i, s_i + 1\}$ by indexing chain-pieces with sequential integers. Therefore, a candidate prior for $\mathcal{S}$ is the distribution induced by simulating a Markov Chain with transition matrix P, whereby only entries $P_{i,i}$ and $P_{i,i+1}$ are non-zero. By simulating $\mathcal{S}$, one can infer $\mathcal{C}$ and K deterministically. A natural prior is as follows. Let $P \sim BetaDiag(a, b)$, where R denotes some maximum number of change-points and $a = a_{1:R}$, $b = b_{1:R}$, if: $P_{i,i} \overset{i.i.d.}{\sim} Beta(a_i, b_i)$, $P_{i,i+1} = 1 - P_{i,i}$, $\forall i \in [1:R]$, $P_{i,j} = 0$, $\forall j \notin \{i, i+1\}$, $P_{R,R} = 1$

Consider the same model setting as in section 2.1, however using the priors described above. This full Bayesian model formulation, first given by Groenewald and Schoeman (2004), for some hyper-parameters a and b, is:

$$\begin{aligned} P &\sim BetaDiag(a, b) \\ \mathcal{S} &\sim MarkovChain(P) \text{ where } s_1 = 1 \\ M_k|\mathcal{S} &\sim DirMat(\boldsymbol{\lambda}) \ \forall k \in [1:K] \end{aligned}$$

It can be seen from equation (2) that a Dirichlet distribution prior on the rows of each transition matrices, M_k, independent across rows, would be conjugate to the likelihood of the observed Markov chain. In the same way, the $BetaDiag$ distribution is also conjugate to the Markov chain of indicators, $\mathcal{S}$.

2.2.1 Implementation and Analysis

Due to the permutations involved with the varying number of change points, it would be very difficult to find the maximum a posteriori (MAP) of the parameters analytically. However, as the model was constructed to be conjugate, it is straight forward to simulate $\mathcal{M}$, P and $\mathcal{S}$ according to the stationary distribution using direct blocked Gibbs sampling Roberts et al. (2004), and hence explore the high density regions of the state-space looking for a mode.

Let $n_k = |\{s \in \mathcal{S} \mid s_j = i, \ \forall j \in [1:N]\}|$ be the number of data-points allocated to chain-piece, k and $z_{i,j}^k = |\{(x_t, x_{t+1}) \in [1,m]^2 | \ s_t = k, \ x_t = i, \ x_{t+1} = j, \ \forall t \in [1:N-1]\}|$ is the number of transitions from state i to state j observed in chain-piece k. Let $a' = a'_{1:R}$ and $b' = b'_{1:(R-1)}$ where $a'_i = a_i + n_i$ and $b'_i = b_i + 1$. $\lambda'^k = \{\lambda'^k_{i,j}\}_{i \in [1:N], j \in [1:N]}$ where $\lambda'^k_{i,j} = \lambda^k_{i,j} + z^k_{i,j}$

The full conditionals of the posterior for Gibbs sampling, in blocks $(P, \mathcal{M}, \mathcal{S})$, is as follows:

$$\begin{aligned} P|x_{1:N}, \mathcal{S} &\sim BetaDiag(a', b') \\ M_k|x_{1:N}, \mathcal{S} &\sim DirMat(\boldsymbol{\lambda'}) \ \forall k \end{aligned}$$

The full conditional distribution for the posterior of $\mathcal{S}$ is trickier to sample from, and less concise to write. However one such way using recursion is detailed fully in Chib (1996) and Groenewald and Schoeman (2004), excluded here in the interest of brevity.

It is also worth noting that the algorithm is very sensitive to choices of a, b in the prior of P. If a is too high then the simulated chain with matrix P may not cycle through sufficient states, and too low then may cycle through states too quickly, and be stuck on the terminating final state.

To further improve the search for the mode, one could use simulated annealing to focus exploration on high density regions of the Markov Chain Monte Carlo (MCMC) state-space. An additional lever to explore would be to use cycles of kernels to navigate the state space more thoroughly, one could simulate $\mathcal{M}$ more often than $\mathcal{S}$ for example. If the number of clusters is proportional to the data length, then this algorithm runs at $\mathcal{O}(n^2)$ complexity per iteration. But due to the nature of MCMC it is difficult to parallelize and many iterations are required. A non-MCMC involving Expectation Maximisation (EM), as suggested by Chib (1996), may perhaps be more practical for finding the mode.

Unlike the next algorithm however, this Bayesian MCMC approach does not introduce any approximation error providing the full state-space can be explored.

2.3 Grid-Merge Heuristic

A very different, but simple bottom-up approach is to split the concatenated chain into chunks, based on fixed width for example, and then recursively merge adjacent chunks based on some threshold and metric of distance between the chain-pieces. As the aim is to group chain-pieces according to the similarity of the latent transition matrix, one could compute the MLE transition matrix, $\hat{M}$ for each chunk then apply a distance metric on the adjacent $\hat{M}$s. The distance metric induced by the Frobenius norm on matrices is a natural choice. For matrices A, B with (i, j) entries, $a_{i,j}, b_{i,j}$ respectively, the distance d is:

$$d(A, B) = \sqrt{\sum_i \sum_j (a_{i,j} - b_{i,j})^2} \tag{4}$$

Given each row of transition matrices are discrete probability distributions, the square of the Frobenius norm is essentially the sum of the squared $L2$ (or squared Wasserstein-2) distance between corresponding distributions. Taking absolute values rather than squared values would be related to the sum of total variation between row distributions, which again is related to the Kullback–Leibler divergence quantity by Pinsker's inequality.

Algorithm 1 Grid-Merge Algorithm

Split chain index $[1 : N]$: $chunks = \{[1 : W], [W + 1 : 2W] \ldots\}$, for some width, W
$Merge \leftarrow TRUE$
while $Merge$ **do**
    $Merge \leftarrow FALSE$
    $j \leftarrow 0$
    for i in odd $Index(chunks)$ **do**
        $j \leftarrow j + 1$
        Compute distance $dist = d(\hat{M}(x_{chunks[i]}), \hat{M}(x_{chunks[i+1]}))$
        if $dist < threshold$ **then**
            $Merge \leftarrow TRUE$
            $new\_chunks[j] \leftarrow merge(chunks[i], chunks[i + 1])$
        else
            $new\_chunks[j] \leftarrow chunks[i]$; $new\_chunks[j + 1] \leftarrow chunks[i + 1]$
            $j \leftarrow j + 1$
    $chunks \leftarrow new\_chunks$
return $chunks$

2.3.1 Implementation and Analysis

The algorithm detailed above may run in $\mathcal{O}(n^2)$ time complexity for the worst case whereby a single merge occurs at each iteration. A less concise algorithm whereby one caches the frequencies at the initial, most refined partition, and not re-run unnecessary comparisons, would run at $\mathcal{O}(n)$.

The initial partition ensures that the maximum number of change points will be $\lfloor N/W \rfloor + 1$, where W is the width of the initial splitting. A large W relative to N and ordered merges would induce a combinatorial limitation, as change-points may only be proposed on entries in the chain index which are multiples of W. Hence, the heuristic may never converge to the "correct" change points, in a frequentist view, only within tolerance upto W. Thus making it important that the width be narrow in hope of getting close to the true change-points. Similarly, in addition to being sensitive to the initial width, the clustering performance of the heuristic is sensitive to the threshold distance chosen. Informally, the choice of initial width and threshold therefore imposes some un-intuitive prior on the output.

For a finite state Markov Chain, Bartlett (1951) showed that the MLE transition matrix, $\hat{M}$ has (i, j) entry: $\hat{m}_{i,j} = \frac{\Omega_{i,j}}{\Omega_i}$, where $\Omega_{i,j} = \sum_{k=1}^{n-1} \delta(X_k = i, x_{k+1} = j)$ and $\Omega_i = \sum_{k=1}^{n-1} \delta(X_k = i)$. Hence, the MLE's behaviour is dependent on the number of visits to each state. It is known, by Pritchard and Scott (2004), that $\forall (i, j) \in [1 : m]^2$, as the sample size grows:

$$\Omega_i^{\frac{1}{2}} (\hat{m}_{i,j} - m_{i,j}) \xrightarrow{D} \mathcal{N}(0, m_{i,j}(1 - m_{i,j})) \tag{5}$$

see Billingsley (1961) for full correlation matrix across transitions. Hence, informally, $\hat{m}_{i,j} \overset{D}{\approx} Z_{i,j}$ where $Z_{i,j} \sim \mathcal{N}(m_{i,j}, \Omega_i^{-\frac{1}{2}} m_{i,j}(1 - m_{i,j}))$

In the merging process given by the heuristic, consider the cleanest case when two adjacent chain pieces are from the same "true" underlying transition matrix with corresponding MLE transition matrices estimates $\hat{A}, \hat{B}$. The square of the Frobenius metric is differentiable, continuous, therefore, using the Continuous Mapping

Theorem, as the size of the adjacent chain-pieces grow:

$$\hat{A}, \hat{B} \xrightarrow{D} A, B, \text{ implies: } d(\hat{A}, \hat{B})^2 \xrightarrow{D} d(A, B)^2 \stackrel{D}{=} \sum_i \sum_j (a_{i,j} - b_{i,j})^2$$

Using equation (5), each $a_{i,j}$, $b_{i,j}$ are correlated normal random variables with constant mean and different variances, hence $d(A, B)$ is distributed according to a generalized Chi-squared distribution, generalized meaning the sum of squared normal random variables, which are not standard. This implies, for this special case, the merging step is, informally, 'similar' to a frequentist Chi-squared hypothesis test and provides some under-developed intuition. The $m_{i,j}$ are however unknown, making transforming the summands in the squared distance metric infeasible, and indeed replacing the unknowns with estimators will then transform the distribution of the squared sum to be perhaps more like a sum of F distributions.

This suggests that this method is suitable for cases whereby the"true" underlying chain-pieces are relatively long, and one can confidently set the initial width sufficiently wide to have convergent, stable transition matrix MLEs. Hence there is a trade-off between a large width causing approximation error, and a narrow width having unstable merge conditions. Further work is required to improve the algorithm in terms of splitting and merging to reduce approximation error and further work required to get a deeper understanding from a probabilistic perspective.

2.4 Empirical Results

Both algorithms are very sensitive to parameter choices. Indeed, for known change-points, such as in this case, it would be straight forward to set an optimum initial width and threshold for the Grid-Merge. However, figure 1 below shows visually that a reasonable, yet imperfect, initial width of 300 and threshold of 0.5 in the Grid-Merge method achieves a visibly good fit, within ∼100 of change points such as index at 1000, 2000 and 8000, due to approximation error. The Bayesian model also shows very close change-points to the true change points, only missing changes points at index 6000 and 8000, which do not appear obvious from the gradient changes in the cumulative counts.

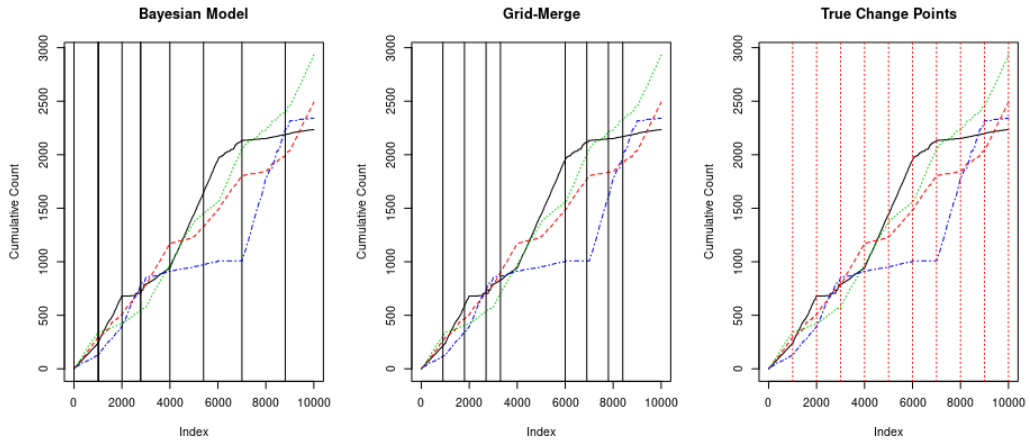

Figure 1: Simulated data: Change points at multiples of 1000 in the index can be seen in the 'kinks' of the cumulative counts of state visits in the 4-state chain. Individual coloured lines are shown for each state

2.5 Final Remarks

Both algorithms show good performance on this relatively small (10,000 observations) chain. However, the algorithms are actually quite complementary. The Grid-Merge heuristic will tend to perform quite well and quickly on large chains with large chain-pieces as the approximation error introduced will be relatively small and the merge criteria relatively stable. The Bayesian approach does not have this approximation error, however, due to requiring many MCMC iterations it would be very slow to converge.

For practical purposes it may be attractive to use the Bayesian model on smaller chains. If the overall chain and true chain-pieces are both large, then first apply the Grid-Merge to break-up the chain, then apply the Bayesian model on the broken up chains independently in parallel to reduce any approximation error. If the chain is large, and the true chain-pieces are small, then perhaps breaking the chain into overlapping chunks, applying the Bayesian method in parallel may be practical.

References

Maurice S Bartlett. The frequency goodness of fit test for probability chains. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 47, pages 86–95. Cambridge University Press, 1951.

Patrick Billingsley. Statistical methods in markov chains. *Ann. Math. Statist.*, 32(1):12–40, 03 1961. doi: 10.1214/aoms/1177705136. URL `https://doi.org/10.1214/aoms/1177705136`.

Siddhartha Chib. Calculating posterior distributions and modal estimates in markov mixture models. *Journal of Econometrics*, 75(1):79–97, 1996.

PCN Groenewald and AC Schoeman. Bayesian detection and analysis of changing transition matrices of stationary markov chains. *Australian & New Zealand Journal of Statistics*, 46(4):555–567, 2004.

Geoffrey Pritchard and David J. Scott. The eigenvalues of the empirical transition matrix of a markov chain. *Journal of Applied Probability*, 41:347–360, 2004. ISSN 00219002. URL `http://www.jstor.org/stable/3215988`.

Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.