The R Package oHMMed - Description and Usage Recommendations

May 12, 2023

Abstract Here, we briefly introduce the R package oHMMed. The first release contains a ready-to-use implementation of a class of ordered Hidden Markov Models with either normal or gamma-poisson emission densities. These are derived and first employed in the following paper: (...tba once available...). Mathematical details of the algorithms, as well as graphical representations of the model structure, can be found there. In this document, we provide a brief overview of our usage recommendations for the algorithms.

Text by Lynette Caitlin Mikula

1 General

In the following, we outline our developers' recommendations for using oHMMed. We consider them best practice pointers, that help ensure that inference is performed as accurately as possible.

Briefly, let us outline the two main assumptions of the oHMMed models and their implications:

(i) The data is an observable sequence of measurements that is generated by a continuously distributed feature with a spatially varying mean–this feature is the 'emission'. Note that these measurements must be taken from non-overlapping windows along the observed sequence, and that there cannot be missing values for any window. The main goal is to assign each window to a hidden/unknown feature 'level' or 'state' (i.e. a discrete category). Subsequently, the parameters determining the distribution of the continuous feature are inferred for each of these states individually. Each window will differ in mean from any window assigned to a different hidden state, but only in stochastic variation from any window assigned the same state. This follows from the assumption on the overall emission density of the feature across the observed sequence. In the case of oHMMed with normal emission densities, this overall emission density should resemble a mixture of bell curves (since each state emits a normal distribution). Note that standard transformations can be applied to the raw data to make its distribution closer to the normal distribution, e.g. to minimise skewness and/or overly light/heavy tails. Obviously, the state-specific parameters inferred by oHMMed will then have to be back-transformed to the original scale for results to be interpretable. In the case of gamma-poisson emission densities, the overall emission density should resemble an overdispersed poisson distribution. Recall that data distributed according to a poisson distributions have equal means and variances; in an overdispersed poisson distribution, the variance should be perceptibly elevated.

(i) Importantly, we assume that the hidden states that generate the observed sequence can be ordered by the increasing means of their individual emission distributions. We further assume that, traversing the hidden sequence, successive windows are only permitted to be assigned to states that are also neighbours within this ordering, *i.e.,* mathematically, the transition rate matrix between the hidden states must be tridiagonal. Overall, this leads to an autocorrelation of the data along the observed sequence, and this should be clearly visible (check, for e.g., by simply plotting). The oHMMed algorithms work well when there are distinct, regional 'blocks' of observed data points with similar values. In other words, the algorithm performs most accurately when the true transition rate matrix between states is highly diagonally dominant (which we will show later). Note that it may be possible to modify the size of the windows along the observed sequences from which the data points are actually calculated to achieve a data resolution that better fulfills this criteria.

2 Setting Priors and Initial Values

Generally, the oHMMed algorithms should be run multiple times with differently specified numbers of hidden states to determine which is most suitable. Particularly with normal emission densities or gamma-poisson emissions with large rate parameters, the number of visible modes, bumps, or perceptible regions of flattening in the gradient of the overall emission density can be an indication of the correct ballpark for this number. However, indications could also come from past literature. This is also true for the prior parameters that specify the state-specific prior distributions. In this case, the priors would be considered strongly informative. Pragmatically, we recommend setting priors using incomplete information from the observed data itself to facilitate quick convergence of the oHMMed algorithms. The restriction of parameters to a feasible range from sneaking a peek at the data is sometimes called regularisation. Priors set in this way are considered weakly informative priors. Note that both the prior distributions and the initial values can be specified by the user in our oHMMed algorithms. If the latter are not explicitly set, they are simply drawn from the prior distributions by the algorithm.

2.1 Normal Emissions

Generally, setting the prior parameters and initial values in the case of normal emission densities is decently straightforward. Since some parameters are fixed behind the scenes (specifically: the coupling constant, and the degrees of freedom), the user only needs to be concerned with setting the means per assumed state and the shared deviation across all states. An additional simplification is that we recommend setting the initial means to the prior means and the initial standard deviation to the prior standard deviation to speed up convergence. This is justified in our parameterisation of the model, since the initial means are then set to the means of the distribution of the prior means and the initial variance is set to a value close to the (technically undefined) mean of the prior distribution of the variance. However, it may be counter-intuitive, since it is often recommended to choose overdispersed initial conditions by sampling the initial values from the prior distribution or distributions. Recall that we implemented this as the default in our algorithms when the initial values are not explicitly set. In this case, however, the MCMC algorithm may need to be run for a longer time to achieve convergence. In terms of how to set the prior means, the approximate locations of the aforementioned irregularities in the observed overall emission density, if present, can be used to inform the choice for the prior values. However, simply remaining within the range of the overall emission density with these and ensuring a decent spacing is generally good enough. Setting the prior standard deviation is potentially more critical—we recommend using a value close to the variance of the full distribution divided by the specified number of states.

We would like to caution the user that, for *e.g.,* too large or small initial values for the standard deviation combined with initial means that fall outside the overall range of the emission distribution could cause the oHMMed algorithm to effectively infer fewer states than specified the user. When this happens, some states simply do not occur in the hidden sequence.

2.2 Gamma-Poisson Emissions

Setting the prior parameters and the initial values in the case of gamma-poisson emission densities is generally not intuitive. The observed sequence is actually a realisation of events drawn from mixture of poisson distributions (and therefore discrete count data), and depends on the rate parameter of the poisson density for each assumed state. This rate parameter is equal to the expected mean and variance of the emission density for each state. However, the algorithm initially emits gamma densities for each state, and hence requires two sets of parameters: There are individual β_i for each assumed state, and a shared α between them; the expected rate parameter in the ith state is the ratio α/β_i. Essentially, the underlying gamma distribution models the spatially varying change in the rate parameter across the observed sequence. Our recommendations for setting initial and prior values extend to cases where the overall emission densities resemble overdispersed poisson densities; other considerations are necessary in cases where the overall emission densities start to resemble a mixture of bell curves. We recommend calculating the mean $\bar{y}$ and the variance s^2 of the data. Then, one can determine the overdispersion with respect to the poisson distribution by subtracting the mean from the variance to obtain the approximate fraction of the variance that requires additional explanation: $s_r^2 = s^2 - \bar{y}$. We then set the initial shape parameter $\alpha^{(0)} = \frac{\bar{y}^2}{s_r^2/K}$ and determine the initial $\beta_i^{(0)}$ to lie within the range of the empirical distribution, and specifically to be smaller than the empirical overall β, which we obtain by dividing $\frac{\sum y}{length(y)}$ by $\frac{\bar{y}^2}{s_r^2}$. Again, we suggest weakly informative priors $\beta_{0i} = \beta_i^{(0)}$ and, noting the similarity of the prior distribution of α to the poisson distribution, $\alpha_{0i} = \alpha^{(0)}$.

2.3 Transition Matrix

In the case of both normal and gamma-poisson emission densities, a prior transition rate matrix between the hidden states must also be specified. Note that both the transition rates between states and the standard deviation of the emitted distributions govern the variability along the observed sequence in our models. For example, frequent transitions between neighbouring states with a emitted distributions with small standard deviations could also be interpreted as one state with a larger standard deviation. Given the interplay between these parameters, success of inference using oHMMed algorithms also depends on the joint initial and prior settings of these parameters. Generally, it is easier to reasonably restrict the range of the initial value and prior distribution of the standard deviation of the emitted distributions (which is directly specified in the normal case and determined by the rate parameter in the gamma-poisson case) than the transition rates. While the initial and prior transition rate matrix can be roughly gauged assuming a set number of states by visually determining how long 'blocks' of similar windows that may be assigned the same hidden state are, we find that the lazy solution of using a randomly generated transition rate matrix is generally sufficient as long as the initial values and the prior distribution of the standard deviation are reasonably specified.

3 Speed and Stability

For all simulations in this section, a total of 600 and 1000 iterations were carried out for the normal and gamma-poisson emissions respectively, with a prespecified 20% burnin period; and convergence was achieved. In our experience with simulated data, thus stipulated performed and discarded iteration numbers have always guaranteed convergence within our recommended framework. In the case of normal emissions, this was sufficient even without setting initial values. For the gamma-poisson emissions, 2000 iterations and a 40% burnin period have generally sufficed in this case. The number of required iteration and length of the burnin period must of course be checked for specific scenarios, specifically on real data. On genomic data, we have used sequences and burnin percentages on the order of twice/quadruple the lengths mentioned here when following our general usage recommendations for the normal/gamma-poisson emission densities respectively.

3.1 Required Sequence Length and Iteration Speed

The accuracy of inference with oHMMed clearly depends on the length of the observed sequence (which we also call the number of windows or L). This is generally given. We recommend running the following simulations after each inference attempt from real data: By simulating an observed sequence using the inferred parameters as the true values, and then trying to re-infer them using an oHMMed run with these true parameters set as priors, one can check the stability of the inference framework for this combination of parameters. We will discuss some general results pertaining to oHMMed's stability here (full inference procedures and results for this section, however, are not shown):

Let us assume three hidden states with transition matrix

$$\begin{bmatrix} 0.990 & 0.010 & 0 \\ 0.005 & 0.99 & 0.005 \\ 0 & 0.010 & 0.990 \end{bmatrix}$$

between them. In a first example, let each state emit a normal distribution with means -1, 0, and 1 respectively, and a shared standard deviation of $\sqrt{0.1}$. Sequences of length greater than the order $L = 2^9$ are required for all hidden states to consistently appear in the simulated data and for the inferred parameters to reach decent accuracy. In a second example, assume that each state emits a gamma-poisson distribution with means $0.2, 1.5$ and 9 respectively (whereby the shared alpha is set to one). Here, sequences longer than order $L = 2^{11}$ are required for all hidden states to appear in the simulated data and the inferred parameters to reach decent accuracy. Note that there is generally less information in an observed sequence with gamma-poisson emission densities than with normal emission densities. Furthermore, the rate parameters most difficult to infer when a gamma-poisson mixture with a low overall rate parameter is emitted are the higher rates (these determine the high-variance 'tails' of the distribution). Performing checks like these for the sequence length and parameter ranges required to ensure there is enough information in the data for the algorithm to work.

Next, we showcase the speed of the oHMMed algorithms for 10 iterations on $L = (2^{11}, 2^{12}, 2^{13}, 2^{14}, 2^{15}, 2^{16})$ windows of the sequences simulated according to the above parameters on a MacBook Pro with an M1 chip running macOS Ventura v13.2.1 (in May 2023). The speed was determined using the R function $Sys.time()$. The time required to complete these iterations increases linearly with sequence length. The slope of this increase is steeper in the case of gamma-poisson emissions: While this version of oHMMed is a touch faster than the version with normal emissions for short sequence lengths, it is noticeably slower for longer sequence lengths. For realistic numbers of iterations, these differences become substantial.

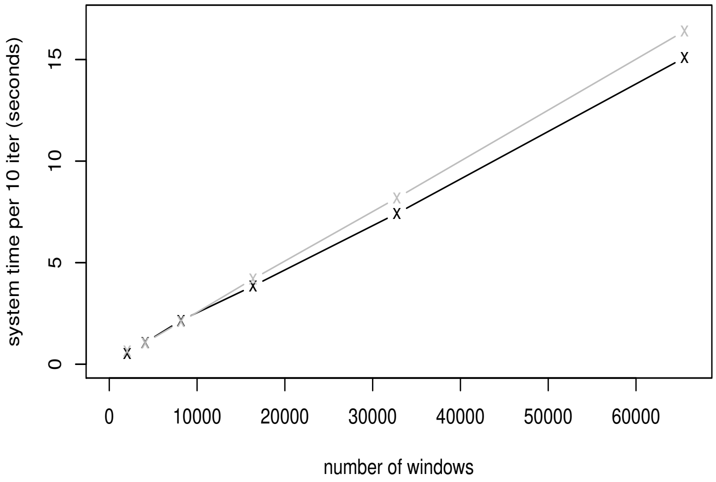

Figure 1: For the simulations described in the main text, the speed of 10 oHMMed iterations is plotted vs. the sequence length, where the 'x' and 'o' mark the speed at lengths $L = (2^{11}, 2^{12}, 2^{13}, 2^{14}, 2^{15}, 2^{16})$ for the case of normal emission densities (in black, approximate values in seconds: 1.067, 2.163, 3.852, 7.423, 15.097), and gamma-poisson emission densities (in grey, approximate values in seconds: 1.052, 2.093, 4.200, 8.187, 16.406) respectively.

3.2 Stability

We have already noted that oHMMed algorithms perform best when the observed sequence exhibits clear 'blocking' of windows that have a similar value for the feature of interest. This occurs when the true transition rate matrix between the hidden states is highly diagonally dominant (provided the standard deviations of the emission densities are not so large relatively that they obscure the blocking effect). Extreme differences in the length of the block of similar windows can cause some states to occur infrequently and in more isolated contexts, which makes it harder for oHMMed algorithms to correctly identify them by their emissions and estimate the values of the emitted densities. Below, we tabulate the effect of different transition rate matrices on inference with the oHMMed algorithms. Note that the version with gamma-poisson emission densities is generally more sensitive to sub-optimal conditions. Again, it is recommended to consider checking these aspects of algorithm stability for the required parameter ranges for each inference attempt with oHMMed.

Table 1: Given the transition matrices in the left column below with the remaining parameters and iteration specifications as in the previous subsection, we simulate a sequence of $L = 2^{13}$ emitted normally distributed data and a sequence of $L = 2^{16}$ emitted gamma-poisson distributed data points. In the right column of the table below, we create confusion matrices of the inferred hidden states (rows) vs the true hidden states (columns). This demonstrates the prediction accuracy of oHMMed for different parameter ranges of transition rates.

True Transition Rates	Confusion Matrix (Normal)	Confusion Matrix (Poisson)
$\begin{bmatrix} 0.990 & 0.010 & 0 \\ 0.005 & 0.99 & 0.005 \\ 0 & 0.010 & 0.990 \end{bmatrix}$	$\begin{bmatrix} 1998 & 2 & 0 \\ 8 & 4639 & 4 \\ 0 & 5 & 1536 \end{bmatrix}$	$\begin{bmatrix} 15371 & 403 & 0 \\ 447 & 33586 & 166 \\ 0 & 318 & 15245 \end{bmatrix}$
$\begin{bmatrix} 0.750 & 0.25 & 0 \\ 0.125 & 0.75 & 0.125 \\ 0 & 0.25 & 0.750 \end{bmatrix}$	$\begin{bmatrix} 1819 & 147 & 0 \\ 121 & 3917 & 126 \\ 0 & 126 & 1936 \end{bmatrix}$	$\begin{bmatrix} 11362 & 4839 & 1 \\ 4417 & 26375 & 2077 \\ 66 & 4392 & 12007 \end{bmatrix}$
$\begin{bmatrix} 0.50 & 0.50 & 0 \\ 0.25 & 0.50 & 0.25 \\ 0 & 0.50 & 0.50 \end{bmatrix}$	$\begin{bmatrix} 1826 & 186 & 0 \\ 174 & 3759 & 181 \\ 0 & 214 & 1852 \end{bmatrix}$	$\begin{bmatrix} 10645 & 5776 & 1 \\ 6123 & 24575 & 2043 \\ 383 & 6142 & 9848 \end{bmatrix}$
$\begin{bmatrix} 0.99 & 0.01 & 0 \\ 0.00125 & 0.99 & 0.00875 \\ 0 & 0.01 & 0.99 \end{bmatrix}$	$\begin{bmatrix} 278 & 0 & 0 \\ 0 & 3547 & 5 \\ 0 & 4 & 4358 \end{bmatrix}$	$\begin{bmatrix} 2820 & 73 & 0 \\ 53 & 32369 & 342 \\ 0 & 402 & 29477 \end{bmatrix}$
$\begin{bmatrix} 0.99 & 0.01 & 0 \\ 0.03125 & 0.75 & 0.21875 \\ 0 & 0.25 & 0.75 \end{bmatrix}$	$\begin{bmatrix} 4968 & 7 & 0 \\ 12 & 1575 & 96 \\ 0 & 95 & 1439 \end{bmatrix}$	$\begin{bmatrix} 40046 & 432 & 0 \\ 643 & 11165 & 1626 \\ 14 & 2928 & 8682 \end{bmatrix}$
$\begin{bmatrix} 0.99 & 0.01 & 0 \\ 0.0625 & 0.50 & 0.4375 \\ 0 & 0.50 & 0.50 \end{bmatrix}$	$\begin{bmatrix} 6098 & 25 & 0 \\ 13 & 1012 & 91 \\ 0 & 66 & 887 \end{bmatrix}$	$\begin{bmatrix} 50818 & 289 & 0 \\ 646 & 6211 & 872 \\ 86 & 2288 & 887 \end{bmatrix}$

4 Diagnostics

Generally, the oHMMed algorithms return two things: Firstly, a sequence of inferred hidden states. Secondly, the following inferred estimates given this sequence of hidden states: the inferred transition rate matrix between them, the inferred parameters of the emission densities per state (these are the means and shared standard deviation for the normal emissions, and the shared alpha, betas, and rates (*i.e.,* alpha through beta) for the gamma-poisson emissions). Convergence behaviour of the oHMMed algorithms as well as the ultimate fit of the inferred model to the observed sequence can be assessed analytically and graphically. In the following, we will outline what to look out for when assessing oHMMed algorithm output. The subsequent section shows specific examples, for which the R-code is also provided and referenced.

4.1 Convergence

As is standard in MCMC diagnostics, the traces (*i.e.,* sequential representation of inferred values across iterations) and the density plots of the inferred parameters and especially the trace of the marginal log-likelihood should be checked for convergence after discarding the burnin period (*i.e.,* a certain number of initial iterations that may be far from the final estimates). Recall that the marginal log-likelihood assesses the overall fit of the inferred model to the observed sequence given the currently assigned hidden states at each iteration. One might expect the marginal log-likelihood to always be negative. However, this is only the case when evaluating discrete emission densities; the sum of probability densities in the case of continuous emissions can, over some (usually small) domains, be positive. Generally, traces of non-problematic MCMC runs are expected to explore the parameter space around a stabilising mean without much serial autocorrelation; and the densities of inferred parameters are meant to be symmetric, unimodal, and decently narrow.

oHMMed algorithms return all the aforementioned traces and densities, and these are visualised as part of the standard plots that are implemented for oHMMed result objects. Behind the scenes, the implementation of these visual diagnostics relies on the R-package *ggmcmc* (Fernández-i-Marín, 2016). These should be used to assess if the length of the burnin was sufficient and if convergence was achieved. For normal emissions, the default settings are 600 iterations with 20% burnin and for gamma-poisson emissions, the default settings are 5000 iterations with 75% burnin.

4.2 Sequence Annotation

Further standard oHMMed diagnostics more explicitly the fit of the inferred model to the observed sequence, and part of these concern the assignment of hidden states to windows along the observed sequence. The number of windows assigned to each state are reported in the standard oHMMed summary function. Standard plotting options include a representation of the observed sequence data, augmented with colouring/shading that shows which hidden state the algorithm has assigned it to. Additionally, the corresponding inferred means are traced along the sequence. Note that these are position-specific posterior means are determined as the sum over the estimated means times the respective probabilities of each state at each position along the sequence. As such, they also contain information on how certain the algorithm is in its assignment of the underlying hidden state. Overall, this plot is a good visual check to see whether there is excess 'mingling' of different states in stretches of the sequence that appear similar to the naked eye, which is an indication that the algorithm has failed to determine distinct distributions for different levels of the emitted feature. If the algorithm is run on simulated data, the same plot is also shown for the sequence of hidden states and true means for a direct comparison. By doing this, the deviations of the inferred posterior means from the true mean become more apparent. Generally, this comparison is particularly informative when viewed alongside the confusion matrix of the true states vs those assigned by the algorithm; this is implemented within the suite of standard oHMMed plots.

4.3 Fit of Overall Distribution

4.3.1 Normal Emissions

Another level at which the model fit can be checked is at the overall emission distribution. The standard oHMMed plot function returns a figure for this: In the case of normal emission densities, the inferred mixture of distributions (where the mixing proportions are determined by the number of windows assigned to each state, and the parameters of the individual normal densities are given by the state-specific inferred parameters) and the observed distribution are overlaid, with the inferred means and 68% confidence intervals (determined from the inferred standard deviation) marked by vertical lines (as are the true means and confidence intervals for runs on simulated data). Emissions emanating from neighbouring states can be distinguished with high statistical confidence when the confidence intervals around their respective means do not overlap. In fact, this is generally a conservative estimate compared to testing for difference in means of the resulting two consecutive states by a one-sided t-test with a 95% confidence interval; the p-values of this are returned by the standard oHMMed summary output. Note that we use the R-function *t.test*() from the library *stats*; if highly significant/in-significant the, p-values are often rounded to 0/1. The oHMMed plot option also returns a classic QQ-plot, which shows the quantiles of the observed distribution vs the quantiles of the inferred mixture of normal distributions. Ideally, the observed quantiles should fall on the diagonal. In order to quantify the deviation of the inferred distribution of the overall emissions from the observed overall distribution, the oHMMed summary function also returns an approximate Kullback-Leibler divergence. We define this as

$$KL_{con}(P,Q) = \sum_{y \epsilon Y^*} p(y)\,\tfrac{p(y)}{q(y)}\,,$$

where P and Q are the probability mass functions and $p(.)$ and $q(.)$ are the probability density functions of the observed and inferred emission data respectively. In our case, $p()$ and $q()$ are computed by the standard *density*() function from the R-package *stats* (R Core Team, 2021); it uses a Gaussian smoothing kernel at 512 equally spaced points; these points make up the set Y^*. The input to the R function *density*() for $p(.)$ and $q(.)$ respectively are the observed feature values and a vector of randomly drawn normally distributed values of the same length as the observed data. These values come from a mixture of normal distributions, whereby the proportions at which each distributions occurs corresponds to the proportion of windows assigned to each state of the feature, and the parameters are set according to the inferred mean and standard deviation of the appropriate state. The output oHMMed returns is the mean of 500 such calculations of $KL_{con}(P,Q)$. A Kullback-Leibler divergence of 0 means that the observed and inferred densities are the same, and values ≥ 0 imply a deviation. Note that because our method is an approximation, values < 0 are possible.

4.3.2 Gamma-Poisson Emissions

Assessment of the fit of the overall distribution inferred by oHMMed is done slightly differently for gamma-poisson emission densities. Again, some methods are graphical and some analytical: The suite of plots for oHMMed algorithm results returns a histogram of the observed feature distribution with the inferred means of each state marked by vertical lines, and the true means in dotted blue vertical lines. In the summary of the output one can find the p-values of exact tests performed on the rate parameters of successive states (in order of increasing mean) using the R-function *poisson.test*() from the library *stats*; if highly significant/in-significant these are often rounded to 0/1. Both of these diagnostics help determine whether emissions from neighbouring states can be statistically distinguished with a high degree of confidence. The oHMMed algorithm also returns a rootogram produced using the R-package *vcd* (Meyer *et al.*, 2022): The observed feature distribution is represented by a barplot. The individual bars are plotted so that their top lines hug the fitted line of the mixture of poisson distributions inferred by oHMMed. Therefore, the amount the bars are shifted from the x-axis shows the extent of the deviation of the fitted model from the observed distribution. Note that the mixture of poisson distributions is generated by averaging over 500 vectors of randomly drawn poisson distributed values of the same length as the observed sequence. Again, these values come from a mixture of poisson distributions, whereby the proportions at which each distributions occurs corresponds to the proportion of windows assigned to each hidden state, and the parameters are set according to the inferred rate of the appropriate state. To put a number on the similarity of the observed and inferred distributions, oHMMed calculates the average Kullback-Leibler divergence of the emissions from 500 sequences generated as just described. The

Kullback-Leibler divergence here is defined as

$$KL(P,Q) = \sum_{y \epsilon Y^*} P(y) \tfrac{P(y)}{Q(y)},$$

where P and Q are the discrete probability distributions of the observed and inferred emission data respectively on the set of individual counts Y^*.

4.4 Optimal Number of States

In the previous 3 subsections, we outlined the methods for assessing individual oHMMed runs for a fixed number of states. Recall that according to our recommended inference framework, multiple runs of oHMMed with differently specified numbers of states should compared. The summary function for oHMMMed will return the mean, median, and standard deviation of the marginal log-likelihood calculated at each iteration from the end of the burn-in period. If the mean and median are similar and the standard deviation small, systematic outliers in the marginal log-likelihood are unlikely and this indicates that convergence of the algorithm is probably good (although this should be checked using the traces as previously discussed). If this is the case, plotting the mean or median log-likelihood against an increasing number of inferred states will yield a visual representation of whether adding more states incrementally increases the model fit or not. Often, there will be a plateau at a certain number of states, after which an increase no longer has as great an effect as before. To corroborate this pattern, one can check whether the (approximate) Kullback-Leibler divergence plateaus in the same way. The number of states at which this levelling of the log-likelihood begins is a good candidate for the optimal choice of states that balances accuracy, generalisability, and interpretability. Regarding the latter, it is also important to check whether the candidate model clearly discriminates between states (using the visual methods and t-tests/ rate tests as previously discussed). If not, revising the choice to a model with fewer states is advisable.

5 Examples

Thus far, we have outlined the rationale for using oHMMed and interpreting its output. In the following, we will demonstrate this rationale on some simulated examples. The R code for these can be found at:

https://github.com/LynetteCaitlin/oHMMed/simulation_scripts

Please run the code alongside the following explanations for a full understanding of how to use oHMMed. We will not reproduce the full results/diagnostics here, and the image quality is not always optimal in this pdf. Details on the functions required to use oHMMed and their specifications (so the arguments used within the functions, for e.g. to set the number of iterations, or to customise plotting options such as deciding to how headers or not,...) can be found on the following website provided by Michal Majka:

https://lynettecaitlin.github.io/oHMMed/reference/index.html

In addition, we provide an outline of the code we use to summarise oHMMed diagnostics for our genetic sequence analyses. This also includes checks of oHMMed assumptions, and assessment of correlations between different genetic sequences that have been separatedly analysed using oHHMed. It can be found in the file oHMMedOutputAnalyses.R alongside the simulation scripts.

5.1 Normal Emissions

5.1.1 Set-up

In this section, we discuss a run of the oHMMed inference framework on simulated data with normal emission densities. Assuming an observed sequence segmented into $L = 2^{13}$ non-overlapping windows, let there be three hidden states with the transitions specified by

$$\begin{bmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.6 & 0.1 \\ 0 & 0.35 & 0.65 \end{bmatrix}.$$

We simulate the normally distributed emissions of these states using means $0.55, 1, 1.9$ and a shared standard deviation of 0.195. Note that with such a transition rate matrix, we expect to incur some mis-assignments of states as part of our inference procedure. Following our advised inference framework, we would run oHMMed on the simulated data with the number of states set to $n = 2, 3, 4, 5$, and setting the initial parameters to the same values as the priors. However, for demonstration purposes we will not set initial values (so these will be drawn from the priors). Per recommendation, we should set the prior standard deviation to the overall standard deviation of the data (which is roughly 0.48) through the number of states in each case. Again purely for demonstration purposes, we set a higher prior standard deviation of 2.5 through the number of states. Further, set the prior means to $0, 3$, and $0.1, 0.8, 3$, and $0.1, 0.7, 0.8, 3$, and $0.1, 0.7, 0.8, 1.5, 3$ respectively; this means some prior means fall outside the range of the data (albeit not grossly so), which is also against our recommendations. However, we stick to our recommendations by using randomly generated prior transition matrices. We let the algorithm run for 600 iterations, and set a 20% burnin.

5.1.2 Results

The mean and median log-likelihoods of this series of runs are plotted for increasing numbers of states in Fig. (2); one clearly sees that they are similar in each case and that they plateau at $n = 3$. The approximate Kullback-Leibler divergence also plateaus here (not shown). This indicates that three states may be the optimal number. We therefore inspect the diagnostics for the run with three hidden states more closely; both the graphical ones using the plot function and the analytical ones using the summary function: Looking at the traces of the log-likelihood and the inferred parameters in Figs. (3,4,5,6), we see that convergence has been reached and the algorithm appears well-behaved. We have omitted the corresponding densities of the inferred parameters here, but these also appear well-behaved. The further diagnostics reveal that the observed and inferred emission densities lie almost directly over each other except for slight deviations in the peaks of the modes. However, the true and observed means and 68-percent confidence intervals appear identical (Fig. (7)). The near-perfect match in overall distribution is reflected in the QQ-plot (Fig. (8)). Note that the confidence intervals of the means do not overlap - hence, emissions from different states can be distinguished with a high statistical certainty (which is confirmed by the highly significant t-test). The exact estimates obtained by the oHMMed runs are $0.5464, 0.9959, 1.8982$ for the means, 0.1990 for the standard deviation, and

$$\begin{bmatrix} 0.7013 & 0.2980 & 0 \\ 0.2859 & 0.6196 & 0.0945 \\ 0 & 0.3215 & 0.6785 \end{bmatrix}$$

for the transition rate matrix. Comparing the assignment of states computed by oHMMed with the true hidden states (Fig. (9), we see that the algorithm separates the two states with the lower means (which are quite close) too harshly: In the simulated sequence, there is a very small amount of 'mingling' of the two states even beyond what could be called the transition zone between them. Recall that if there is complete 'mingling' of the $inferred$ states, the oHMMed run has failed to distinguish them (either because of poor performance or true 'mingling' in the observed sequence). The confusion matrix reflects both the appreciable accuracy and the visible mis-specifications:

$$\begin{bmatrix} 3195 & 458 & 0 \\ 282 & 3179 & 8 \\ 0 & 11 & 1059 \end{bmatrix}.$$

Overall, this series of runs performs as one would hope.

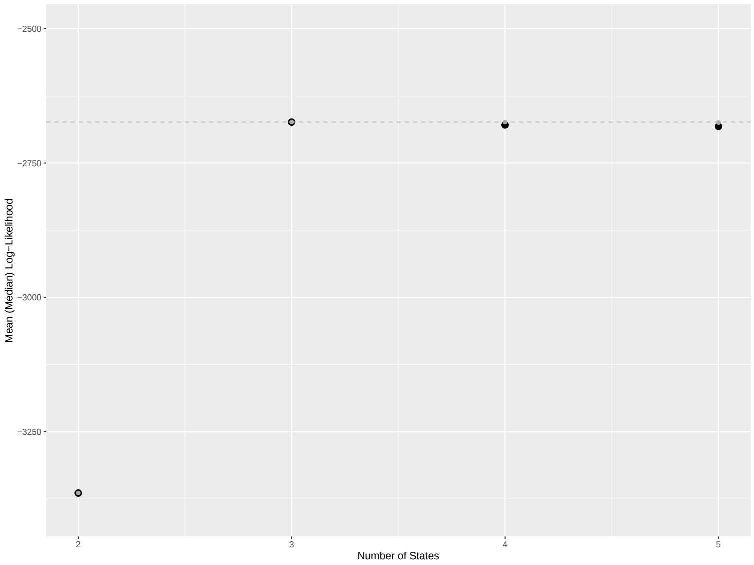

Figure 2: The filled circles in these panels demark the mean and median log-likelihoods (in black and grey respectively) for the inference procedures on the simulated data described in the above text. The dotted horizontal line represents the point halfway between the mean and median log-likelihood of the 'chosen' number of states.

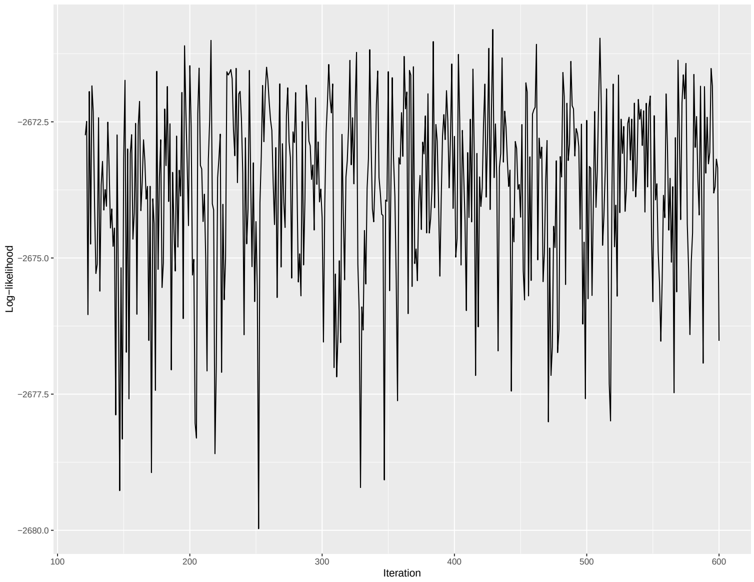

Figure 3: Trace of the log-likelihood, after discarding the burnin.

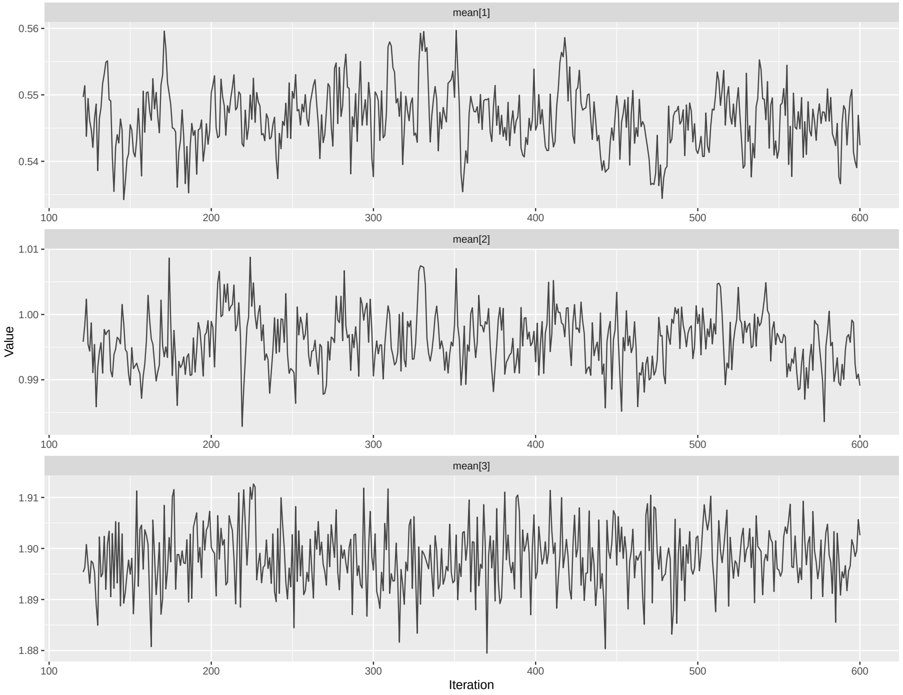

Figure 4: Trace of the individual inferred means, after discarding the burnin.

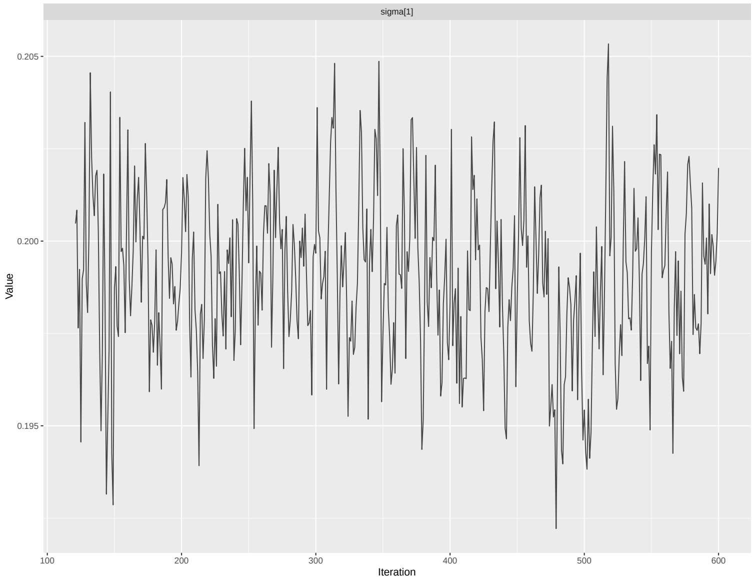

Figure 5: Trace of the inferred standard deviation, after discarding the burnin.

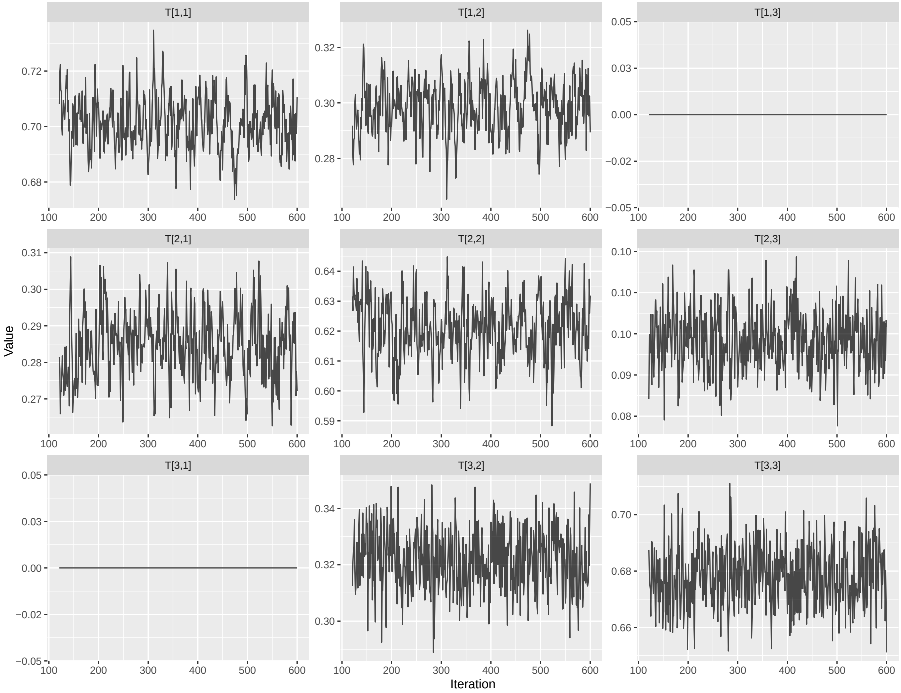

Figure 6: Traces of the inferred entries of the transition rate matrix, after discarding the burnin.

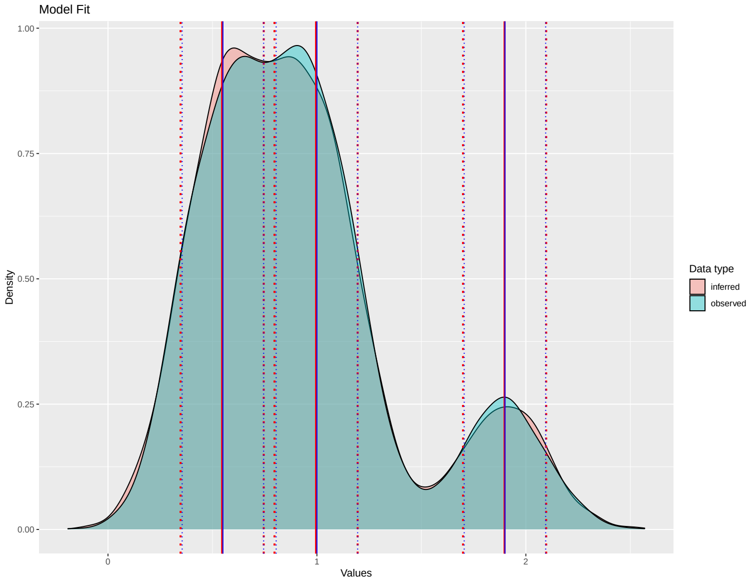

Figure 7: The observed and inferred overall emission densities are over-layed; the individual densities are in red and turquoise respectively. Solid vertical lines represent the observed and inferred means according to the same colour scheme, and the bounds of the 68-percent confidence intervals per inferred mean are marked by dotted vertical lines.

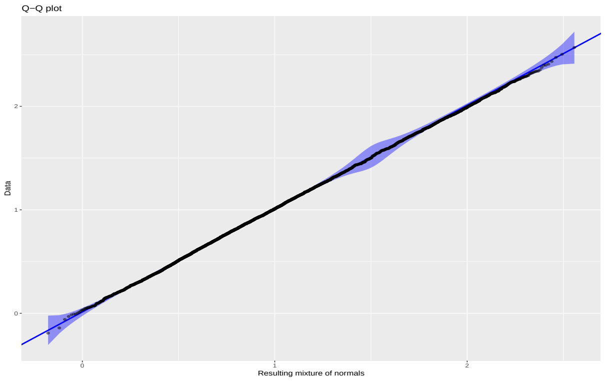

Figure 8: QQ-plot of the inferred vs the observed densities from the previous figure.

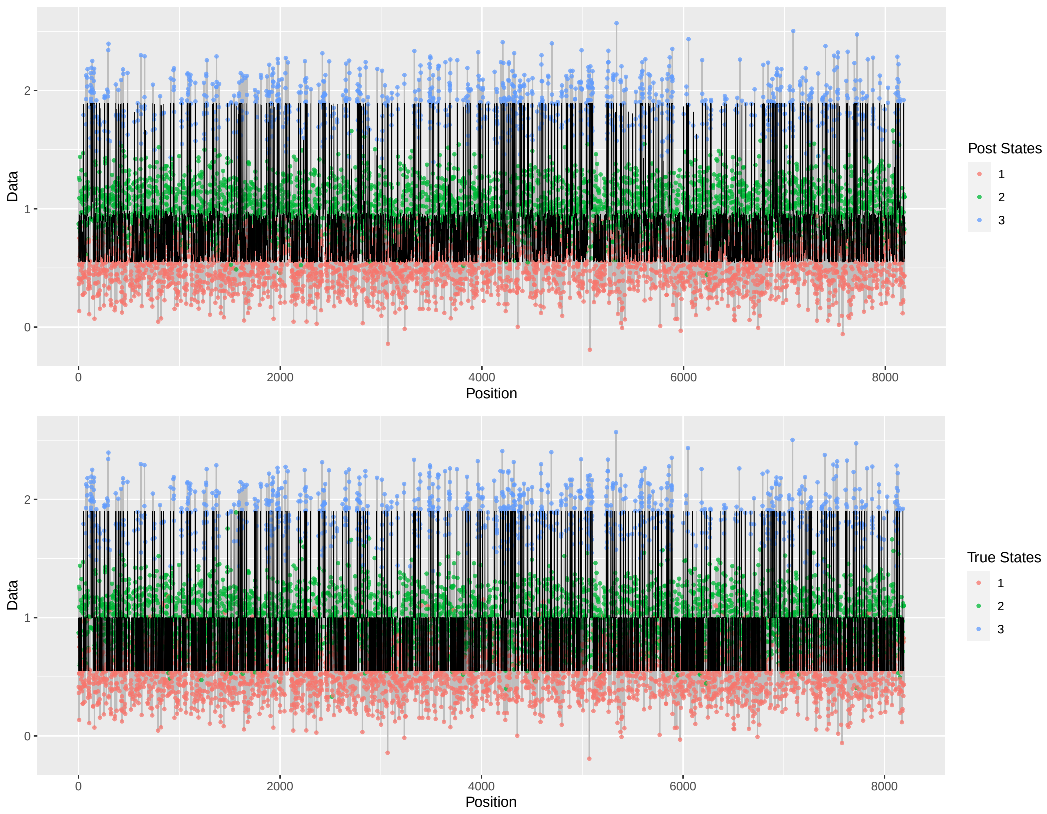

Figure 9: These plots show the data points/emissions for each segment along the sequence coloured by the hidden states. In the upper plot, these are the true hidden mean; in the lower plot, these are the inferred hidden states. The black lines trace the inferred posterior mean across the sequence. Recall that these position-specific posterior means are the sum of estimated means times the respective probabilities of each state.

5.2 Gamma-Poisson Emissions

5.2.1 Set-up

In this section, we document an example of an oHMMed inference run on simulated data with gamma-poisson emission densities. Assuming an observed sequence segmented into $L = 2^{13}$ non-overlapping windows, let there be three hidden states with the transitions specified by

$$\begin{bmatrix} 0.99 & 0.01 & 0 \\ 0.01 & 0.98 & 0.01 \\ 0 & 0.01 & 0.99 \end{bmatrix} .$$

We simulate the gamma-poisson distributed emissions of these states with an alpha of 1.3 and individual betas of 5, 0.66, 0.11. Note that with such a transition rate matrix, we expect inference to be fairly accurate. Then, we run oHMMed on the simulated data with the number of states set to $n = (2, 3, 4, 5)$. For demonstration purposes, we disregard our own recommendations in that we do not set initial values; these are then drawn by default from the priors. The prior alphas are set to the recommended priors, and prior betas to $5, 1$, and $5, 3, 1$, $5, 4, 3, 1$, and $5, 4, 3, 2, 1$ respectively (also in keeping with recommendations). We let the algorithm run for 2000 iterations, and set a 40% burnin.

5.2.2 Results

The mean and median log-likelihoods of the resulting oHMMed models are plotted for increasing numbers of states in Fig. (10); one clearly sees that they are similar in each case and that they plateau at $n = 3$. The approximate Kullback-Leibler divergence also plateaus here (not shown). This indicates that three states may be the optimal number, and looking at the trace and density plots of the log-likelihood and the inferred parameters (Figs. (11), (12), (13), (14)) for this particular run, we see that convergence has been reached and the algorithm appears decently well-behaved in spite of some autocorrelation in some traces (which would be improved by performing more iterations). Note that we have omitted the corresponding density plots for the inferred parameters, as well as the trace and density plots for the means of each state.

The exact estimates obtained by the oHMMed runs are 1.343 for alpha, 5.119, 0.665, 0.111 for the betas, and

$$\begin{bmatrix} 0.9895 & 0.0105 & 0 \\ 0.0105 & 0.9817 & 0.0078 \\ 0 & 0.0111 & 0.9888 \end{bmatrix}$$

for the transition rate matrix. This immediately shows good performance. Note that the estimates of the means per state can often be more accurate than the individual estimates for alpha and the betas themselves; we do not report the exact values here but show them in Fig. (15, where we see that inferred means are accurate - only the last one is slightly underestimated. The inferred rate parameters/means of neighbouring states are also significantly different (exact test printed in the summary but not shown here). The fit of the overall inferred distribution of gamma-poisson mixtures is generally good (Fig. (16)) except for difficulties with the high variability and scarcity of data in the tail of the overall distribution, which explains the decreased accuracy of the estimators in this region. Comparing the assignment of states computed by oHMMed with the true hidden states (see Fig. (17) - note that oHMMed also plots these on a log scale if this is preferred for better visibility), we see that it is very accurate. Recall that if there is complete 'mingling' of the $inferred$ states, the oHMMed run has failed to distinguish the two (either because of poor performance or true 'mingling' in the observed sequence). The confusion matrix reflects the appreciable accuracy, and reveals a few mis-specifications:

$$\begin{bmatrix} 2956 & 85 & 0 \\ 58 & 2942 & 32 \\ 0 & 16 & 2103 \end{bmatrix}.$$

Overall, this series of runs performs as one would hope.

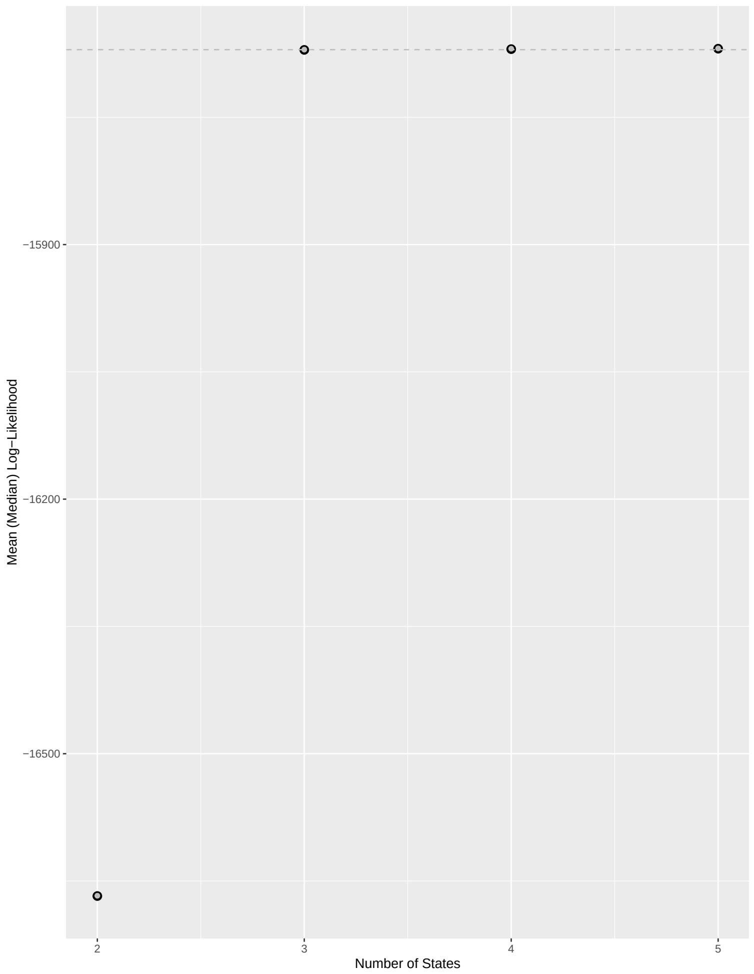

Figure 10: The filled circles in these panels demark the mean and median log-likelihoods (in black and grey respectively) for the inference procedures on the simulated data described in the above text. The dotted horizontal line represents the point halfway between the mean and median log-likelihood of the 'chosen' number of states.

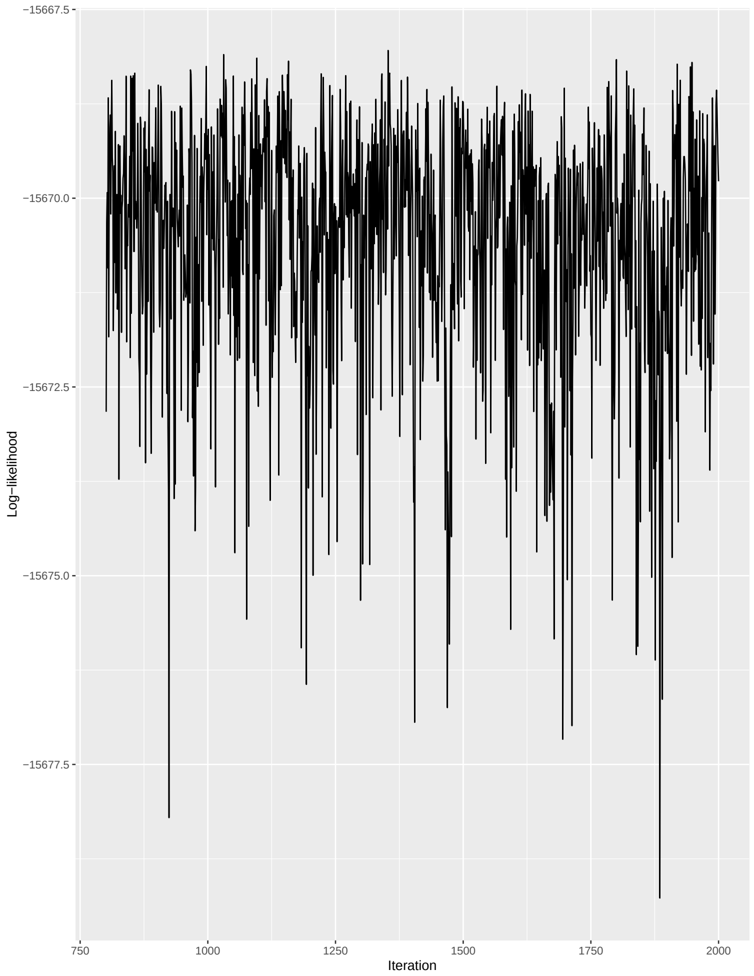

Figure 11: Trace of the log-likelihood, after discarding the burnin.

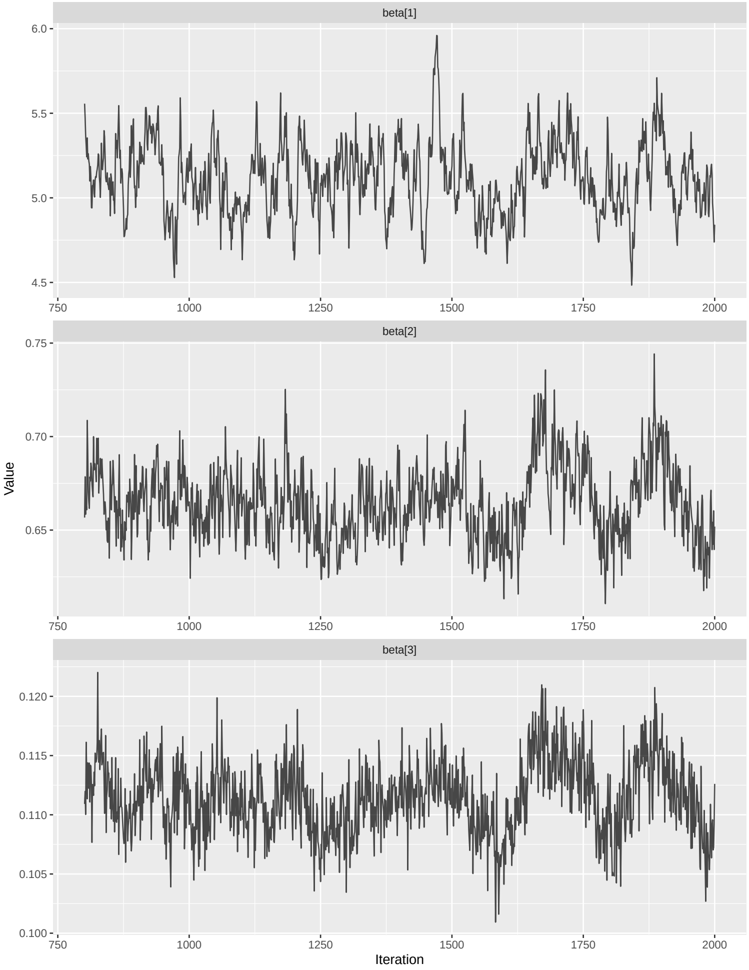

Figure 12: Trace of the individual inferred betas, after discarding the burnin.

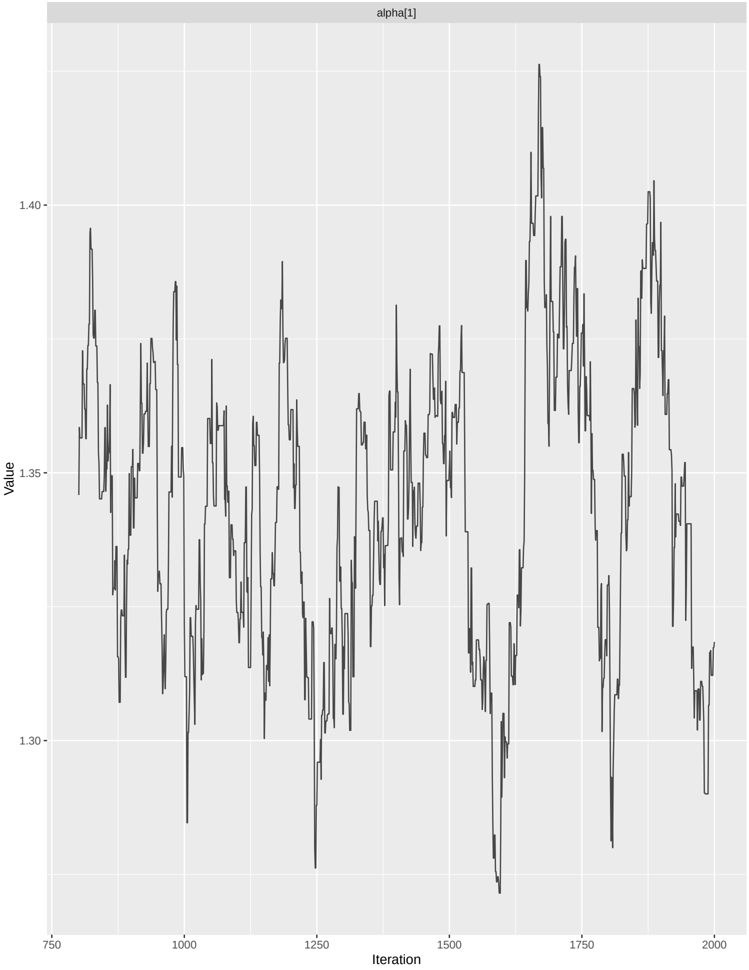

Figure 13: Trace of the inferred alpha, after discarding the burnin.

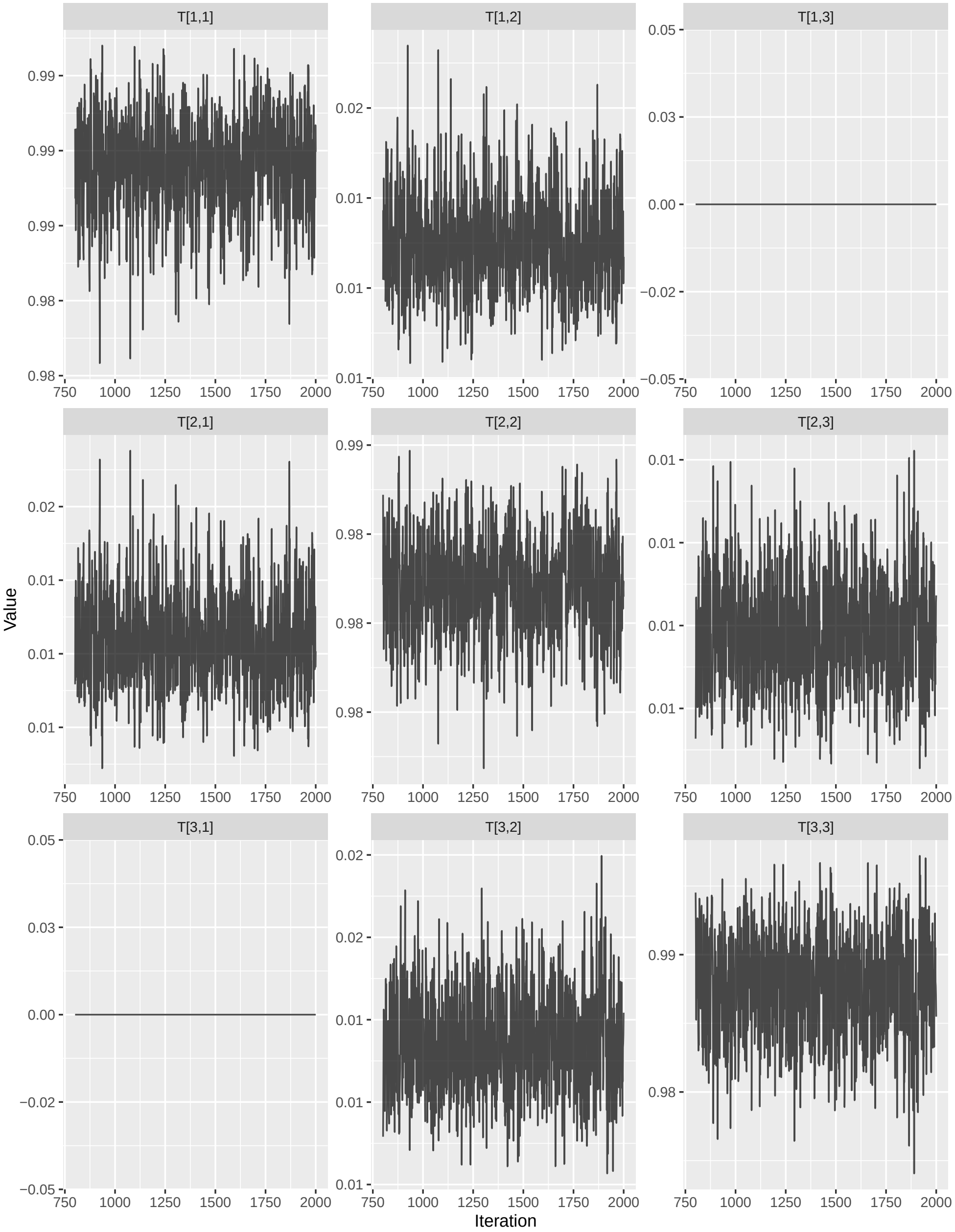

Figure 14: Traces of the inferred entries of the transition rate matrix, after discarding the burnin.

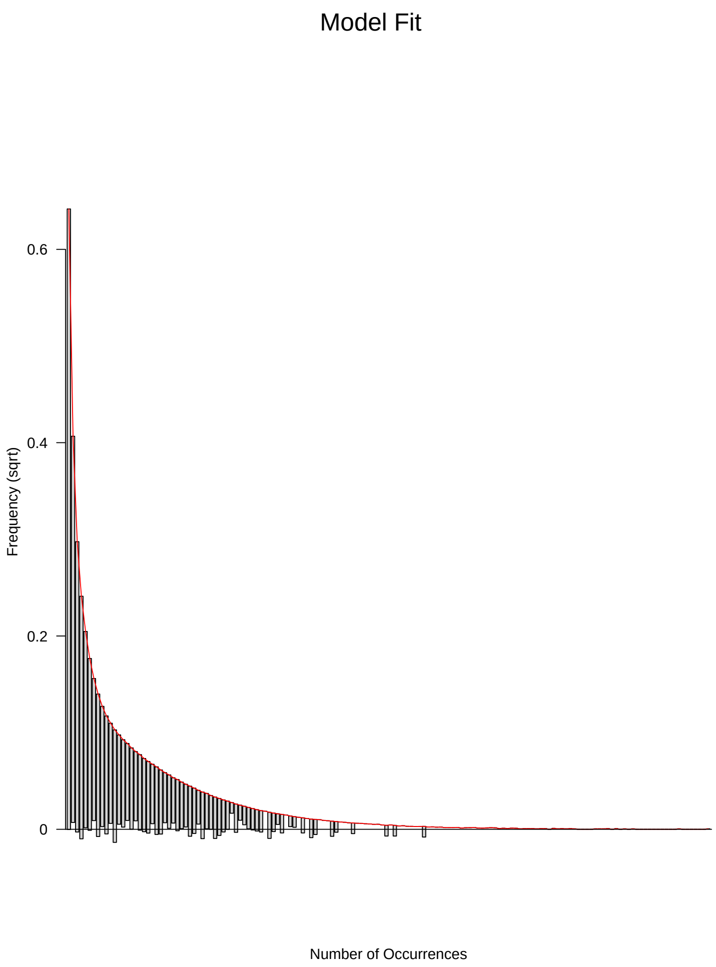

Figure 15: Rootogram of the observed data (vertical bars) fitted to the inferred mixture of poisson distributions (smoothed red line).

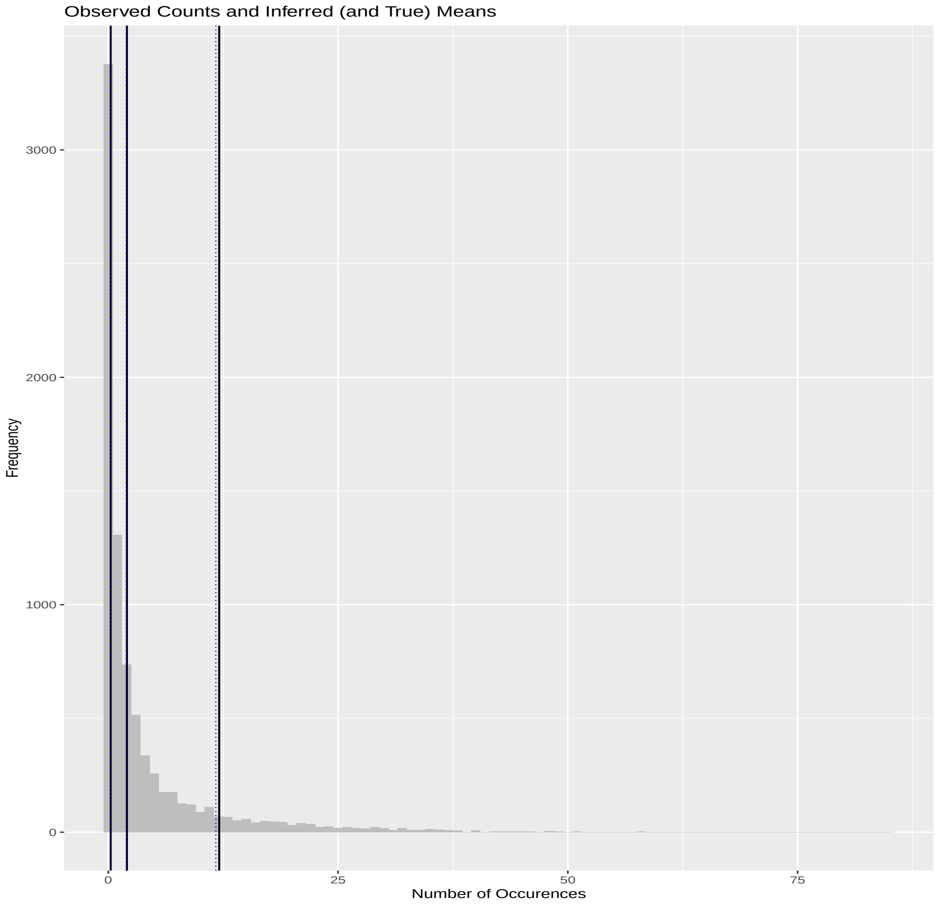

Figure 16: Histogram of the observed data with the true means in the solid vertical lines, and the inferred means as dotted vertical means.

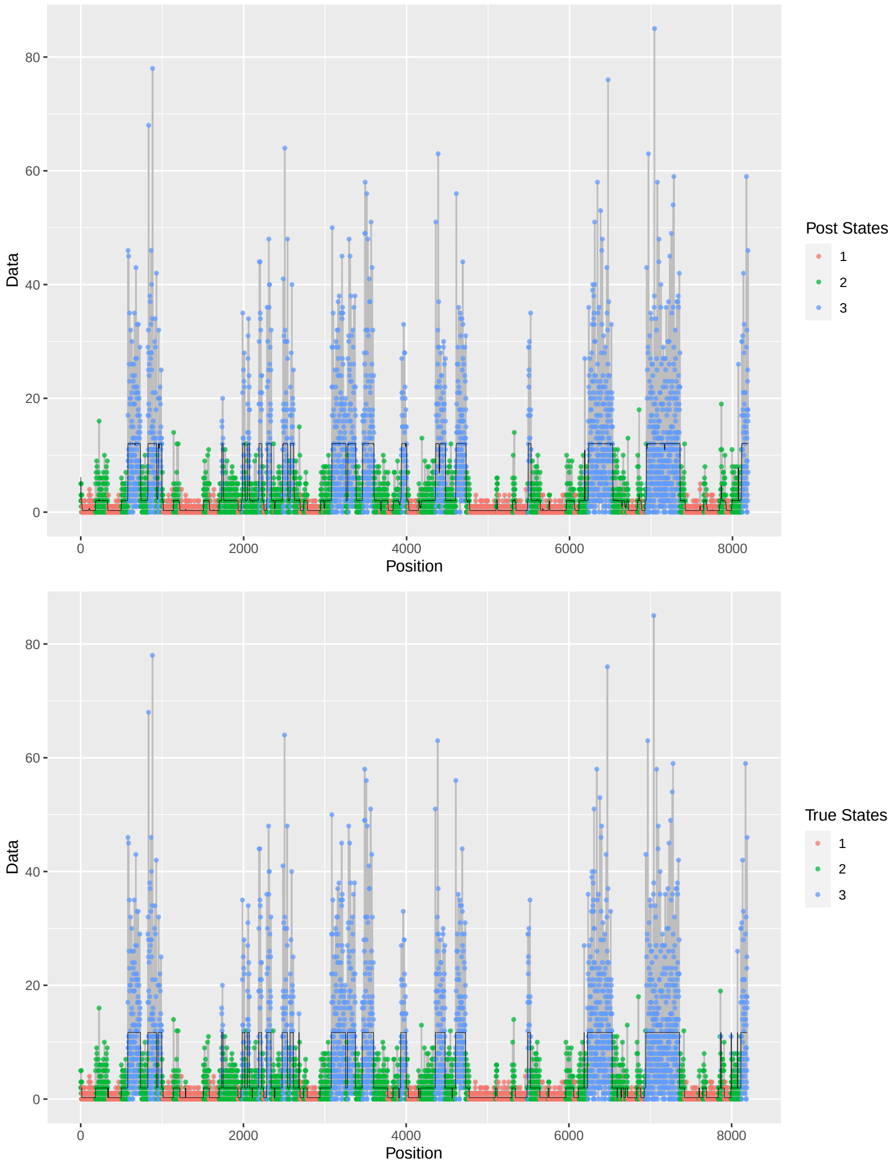

Figure 17: These plots show the data points/emissions for each segment along the sequence coloured by the hidden states. In the upper plot, these are the true hidden mean; in the lower plot, these are the inferred hidden states. The black lines trace the inferred posterior mean across the sequence. Recall that these position-specific posterior means are the sum of estimated means times the respective probabilities of each state.

References

Fernández-i-Marín, X. (2016). ggmcmc: Analysis of MCMC Samples and Bayesian Inference. *Journal of Statistical Software*, **70**(9), 1–20.

Meyer, D., Zeileis, A., and Hornik, K. (2022). *vcd: Visualizing Categorical Data. R package version 1.4-10.*

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.