The R Package oHMMed - Description and Usage Recommendations

March 18, 2023

Here, we briefly introduce the R package oHMMed. It contains a ready-to-use implementation of the ordered Hidden Markov Models with normal and gamma-poisson emission densities derived and employed in the following paper: (...tba once available...). Mathematical details of the algorithms, as well as graphical representations of them, can be found in this paper. Below, we provide a brief overview of our usage recommendations.

1 General

In the following, we outline the best practice for employing oHMMed with normal and gamma-poisson emission densities. Let us outline the main assumptions of the model:

The first is that there is an observable sequence of data that is shaped by a continuously distributed feature with a spatially varying mean–this feature is the 'emission'. Assuming that the value of this feature has been determined for non-overlapping windows along the sequence, our main goal is to assign each window to a hidden/unknown feature 'level' or 'state' and infer the parameters of the feature distribution for these states: Each window will differ in mean from windows assigned a different state, but only in stochastic variation from windows assigned the same state. In the case of normal emission densities, the overall distribution of the feature across the observed sequence should thus resemble a mixture of bell curves. Note that standard transformations can be applied to the raw data to minimise skewness and/or overly light/heavy tails to make this model assumption more applicable–the parameters inferred by the algorithm for the distributions of the feature per state then have to be back-transformed for correct interpretation. In the case of gamma-poisson emission densities, the overall distribution of the feature across the observed sequence should resemble an overdispersed poisson distribution; so while poisson distributions have equal mean and variance, the variance of the feature should be perceptibly increased.

The second main assumption is that the hidden states can be ordered by the increasing means of their emission distributions: Traversing the hidden sequence, successive windows are only permitted to be assigned to states that are

also neighbours within this ordering, *i.e.,* the transition rate matrix between the hidden states must be tridiagonal. This autocorrelation should be visible when plotting the observed sequence; the algorithm works well when there are distinct 'blocks' of windows with similar means and levels of variability in terms of the focal feature. 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 sequences to achieve a 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 distribution of the feature across the entire observable sequence 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 distributions of the variables of interest. Such priors are known as strongly informative priors. Pragmatically, we recommend setting priors using incomplete information from the data itself to facilitate quick convergence of the algorithms. This restriction of parameters to a feasible range is sometimes called regularisation; such priors are known as weakly informative priors. Note that both the prior distributions and the initial values can be specified by the user in the oHMMed algorithms. If the latter are not explicitly set, they are simply drawn from the prior distributions.

2.1 Normal Emissions

Generally, setting the priors and initial values in the case of normal emission densities is decently straightforward. Since some parameters are fixed behind the scenes (the coupling constant κ_0 and the prior degrees of freedom), the user only needs to be concerned with the means per assumed state and the shared deviation across all states: Further, we recommend setting the initial means to the prior means and the initial standard deviation to prior standard deviation to speed up convergence. This is justified using our parameterisation since the initial means are then set to the means of the distribution of prior means and the initial variance is 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 long time to converge. In terms of how to set the initial and prior means, the approximate

locations of the aforementioned irregularities in the observed overall density, if present, can be used to inform the choice. However, simply remaining within the range of the overall distribution and ensuring a decent spacing is generally good enough. Setting the initial and 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. For *e.g.,* too large initial values or too small initial values combined with initial means falling outside the overall range of the distribution could cause the algorithm to effectively infer fewer states than specified by not assigning any parts of the hidden sequence to one or more of the hidden states.

2.2 Gamma-Poisson Emissions

Setting the initial values and prior parameters in the case of gamma-poisson emission densities is less intuitive: The observed sequence is actually a mixture of poisson distributions, so discrete count data, and depends on the rate parameters per assumed state. This rate is also the expected mean value and variance of the emitted count data for each state. However, the algorithm itself emits gamma densities 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, this underlying gamma distribution models the spatially varying change in the rate parameter across the observed sequence. Our recommendations for 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 is to be explained by the process: $s_r^2 = s^2 - \bar{y}$. 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, $\upsilon_0 = \alpha^{(0)}$.

2.3 Transition Matrix

In the case of both normal and gamma-poisson emission densities, a prior transition rate matrix 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 that share a small standard deviation could also be interpreted as one state with a larger standard deviation. Given the interplay between these parameters, success of inference

using oHMMed also depends on the joint 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 feature (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 specific number of states by visually determining how long 'blocks' of similar windows that may be assigned the same hidden state are, we find that a 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. All algorithms converged. In our experience, thus stipulated performed and discarded iteration numbers have always guaranteed convergence within our recommended framework and even without setting initial values in the case of normal emissions. For the gamma-poisson emissions, 2000 iterations and a 40% burnin period have always sufficed under the same conditions. The number of required iteration and length of the burnin period must of course be checked for specific scenarios.

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 call the number of windows or L). We recommend running simulations after each inference attempt: By simulating an observed sequence using the inferred parameters as the true parameters, 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 the required combination of parameters.

In order to gauge oHMMed's performance, we discuss some examples here (full results, 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 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). Checks such as this should always be performed for the sequence length and parameter ranges required to ensure there is enough information in the data for the algorithm to work. After observing this, we check the speed of oHMMed for 10 iterations on $L = (2^{12}, 2^{13}, 2^{14}, 2^{15})$ windows of the sequences simulated according to the above parameters on a MacBook Pro with an M1 chip running macOS Monterey v12.4. The speed was determined using the R function $Sys.time()$. The time required to complete these iterations increases linearly with sequence length. The algorithm with normal emission densities is a touch faster than the algorithm with gamma-poisson emission densities; for realistic numbers of iterations this becomes a palpable difference.

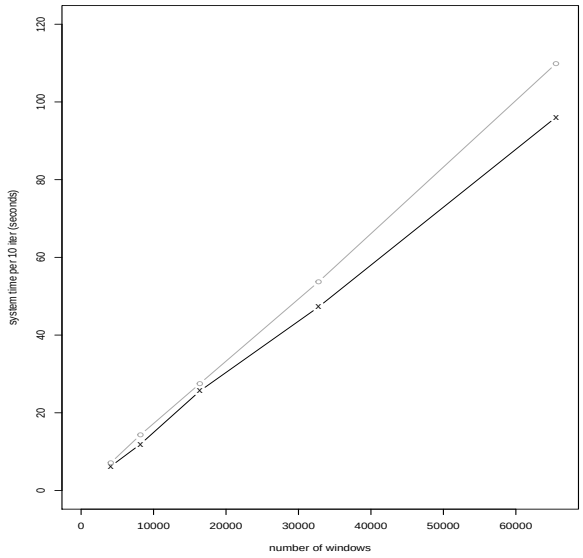

Figure 1: For the simulations described in the main text, the speed of 10 oHMMed iterations is plotted vs. the sequence length, where the 'x' and 'o' mark the lengths $L = (2^{12}, 2^{13}, 2^{14}, 2^{15})$ for the case of normal emission densities (in black) and gamma-poisson emission densities (in grey) respectively.

3.2 Stability

We previously noted that oHMMed performs best when the observed data exhibits clear 'blocking' of windows that are similar in the focal feature. This occurs when the true transition rate matrix between the true hidden states is highly diagonally dominant (provided the shared standard deviations are not so large relatively that they obscure the borders between blocks). Extreme differences in the length of these blocks can cause some states to occur infrequently and in more isolated contexts, which makes it harder to correctly identify them by their emissions. Below, we tabulate the effect of different transition rate matrices on inference using oHMMed. Note that once again, effects are more pronounced with gamma-poisson emission densities. Again, it is recommended to check this as part of every oHMMed inference run.

Table 1: Given the transition matrices in the left column below and the remaining parameters and iteration specifications as in the above texts, we simulate a sequence of $L = 2^{13}$ emitted normally distributed 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 states (rows) vs true states (columns). This demonstrates the prediction accuracy of oHMMed for different parameter ranges.

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

The oHMMed algorithms return a sequence that assigns each segment of the hidden sequence to a state (the number of segments assigned to each state are also tabulated in the summary of the output for a quick overview). Furthermore, they return the inferred entries of the transition rate matrix between the hidden states, and 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 both analytically and graphically.

4.1 Convergence

As is standard in MCMC diagnostics, the traces (*i.e.,* sequential values for each iteration) and the density plots of the inferred parameters as well as 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 fit of the observed values of the feature along all windows given the assigned states at each iteration. One might expect the marginal log-likelihood to always be negative, but this is only the case when evaluating discrete emissions. However, the sum of probability densities in the case of continuous emissions can, over some (usually small) domains, be positive. All of the aforementioned traces and densities are visualised as part of oHMMed output plots and one must here assess if the length of the burnin was sufficient; the default is set to 20% of the total sequence length/number of windows/L. Behind the scenes, the implementation of these visual diagnostics relies on the R-package *ggmcmc* (Fernández-i-Marín, 2016). 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.

4.2 Sequence Annotation

Further diagnostics concern the fit of the inferred values for the observed sequence. The oHMMed algorithms return a sequence level representation of the results: A plot of the observed sequence is augmented with colouring/shading that shows which hidden state the algorithm has assigned it to, and 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. 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 de-

termine distinct distributions for different levels of the emitted feature. If the algorithm is run on simulated data, the same plot is shown for the true states and true means for a direct comparison. This is particularly informative when viewed alongside the confusion matrix of the true states vs those assigned by the algorithm; this is also returned by oHMMed.

4.3 Fit of Overall Distribution

4.3.1 Normal Emissions

Another level at which the model fit can be checked is that of the overall distribution of the feature: 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 respective 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. oHMMed 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, oHMMed 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 oHMMed algorithm 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 state of the feature, 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

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

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

We have now 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. Each run 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. If not, revising the choice to a model with fewer states is advisable.

5 Examples

The R code for the following examples can be found at:

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

Please use these alongside the following explanations. We will not reproduce the full results/diagnostics here, and the image quality is not always optimal.

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 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: 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. 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.

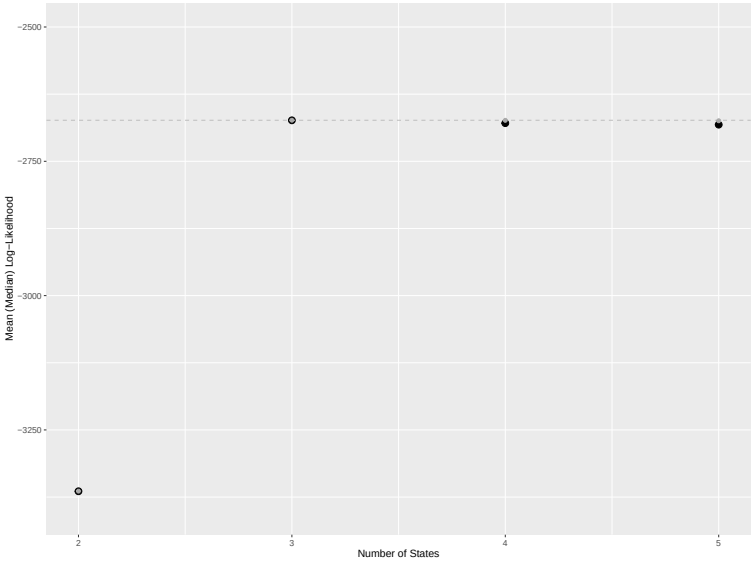

Figure 2: The filled circles in these panels demark the mean and median log-likelihoods (in black and grey respectively) for the inference procedures on the simulated data described in the above text. The dotted horizontal line represents the point halfway between the mean and median log-likelihood of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

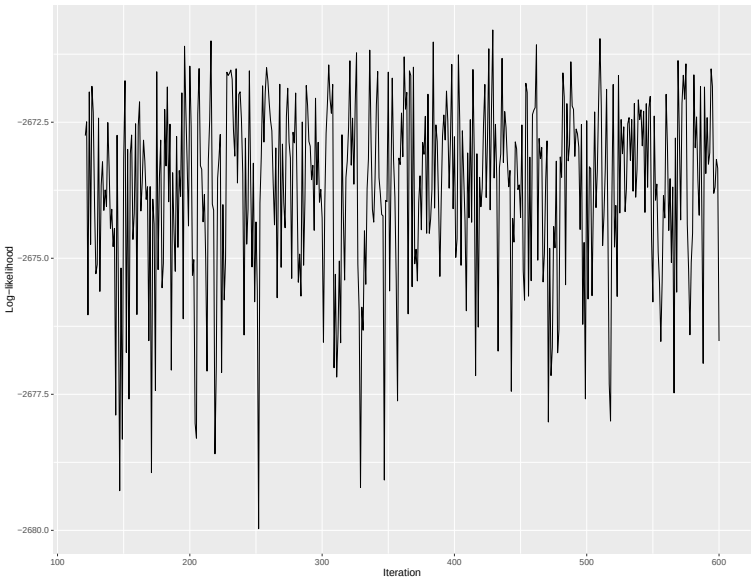

Figure 3: Trace of the log-likelihood, after discarding the burnin.

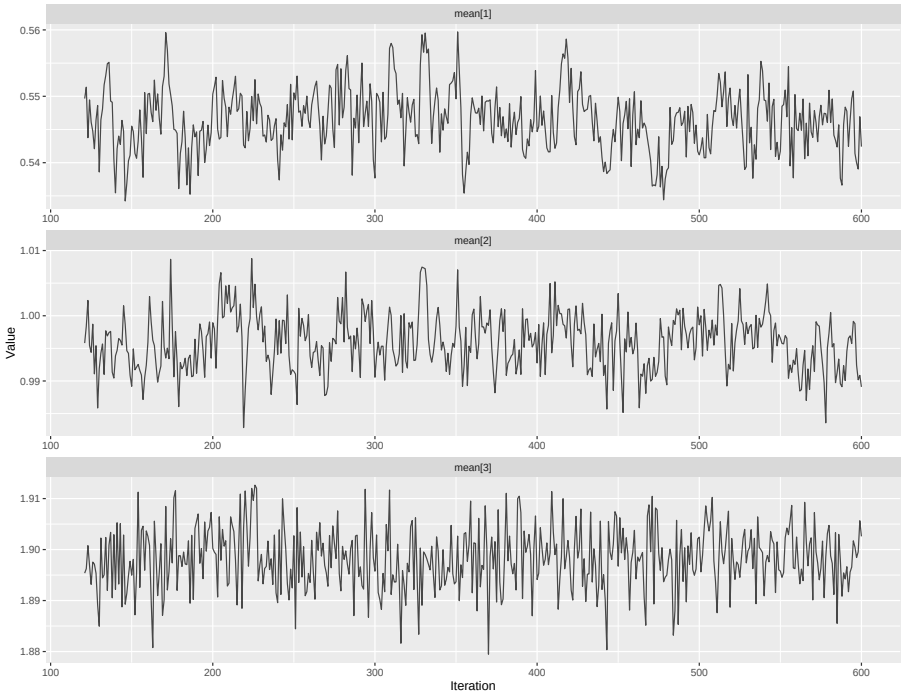

Figure 4: Trace of the individual inferred means, after discarding the burnin.

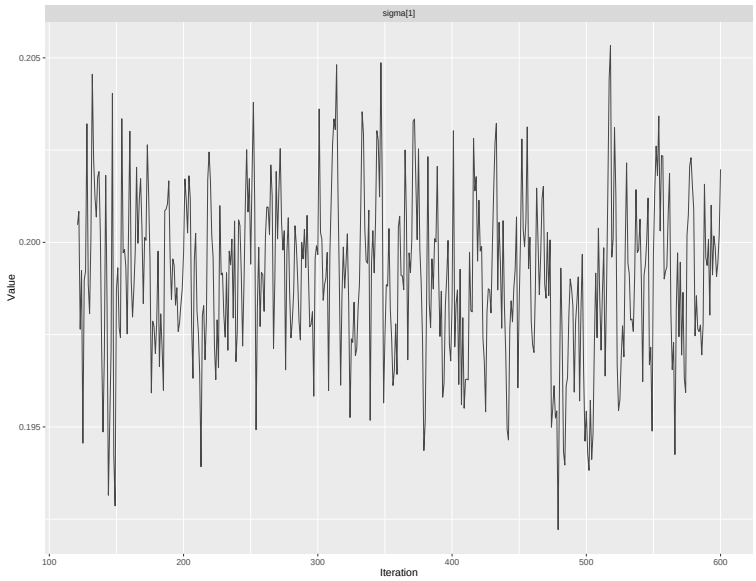

Figure 5: Trace of the inferred standard deviation, after discarding the burnin.

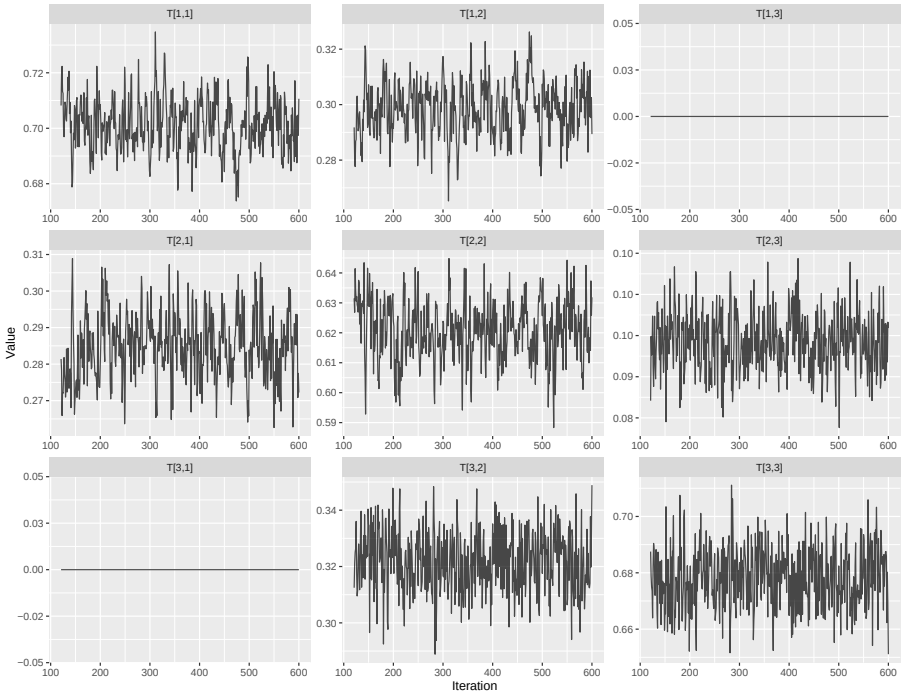

Figure 6: Traces of the inferred entries of the transition rate matrix, after discarding the burnin.

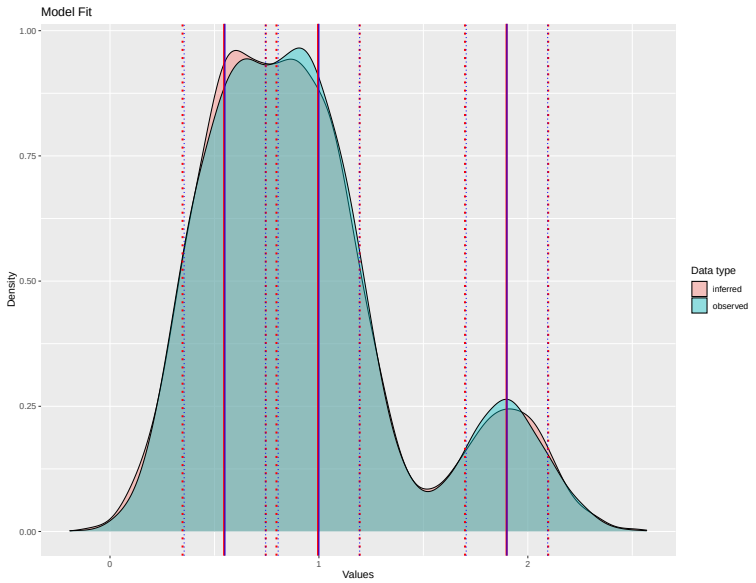

Figure 7: The observed and inferred overall emission densities are over-layed; the individual densities are in red and turquoise respectively. Solid vertical lines represent the observed and inferred means according to the same colour scheme, and the bounds of the 68-percent confidence intervals per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

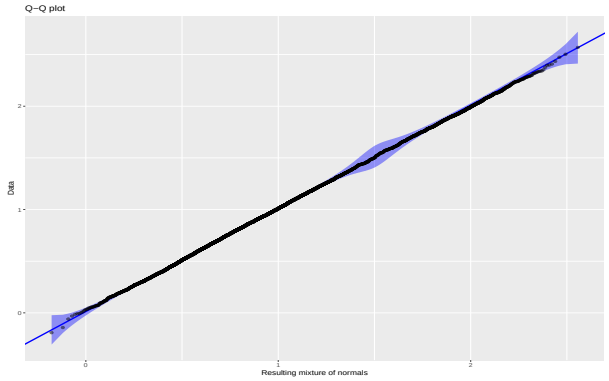

Figure 8: QQ-plot of the inferred vs the observed densities from the previous figure.

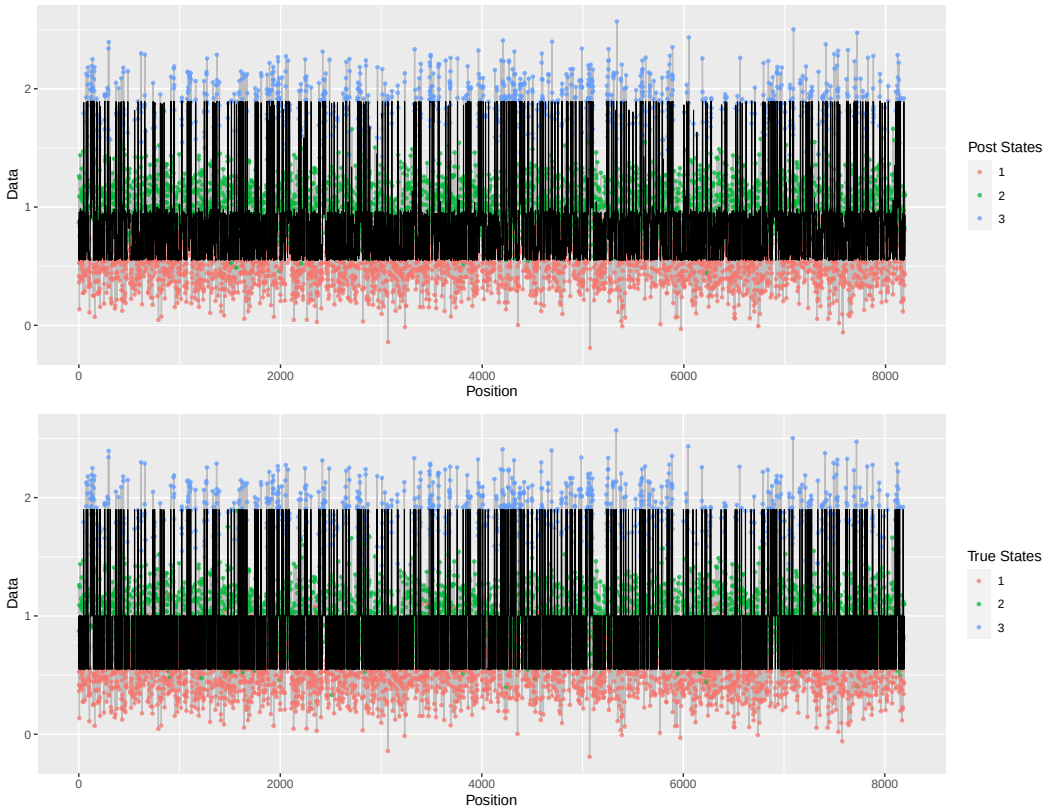

Figure 9: These plots show the data points/emissions for each segment along the sequence coloured by the hidden states. In the upper plot, these are the true hidden mean; in the lower plot, these are the inferred hidden states. The black lines trace the inferred posterior mean across the sequence. Recall 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 poisson-gamma 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. 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 (Fig. (17)), 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.

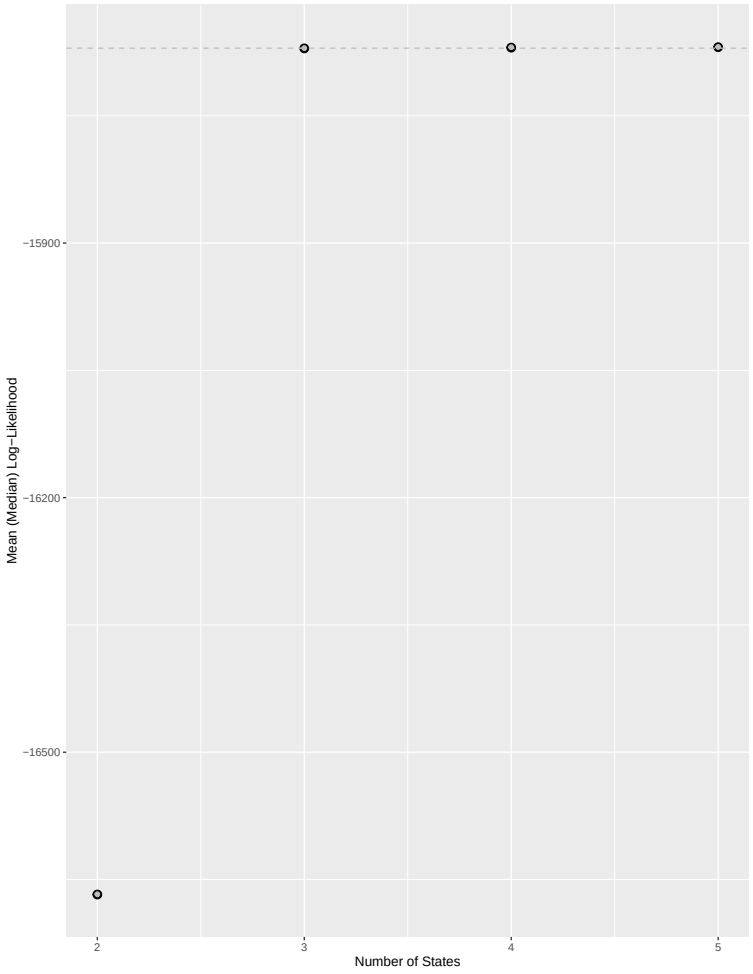

Figure 10: The filled circles in these panels demark the mean and median log-likelihoods (in black and grey respectively) for the inference procedures on the simulated data described in the above text. The dotted horizontal line represents the point halfway between the mean and median log-likelihood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

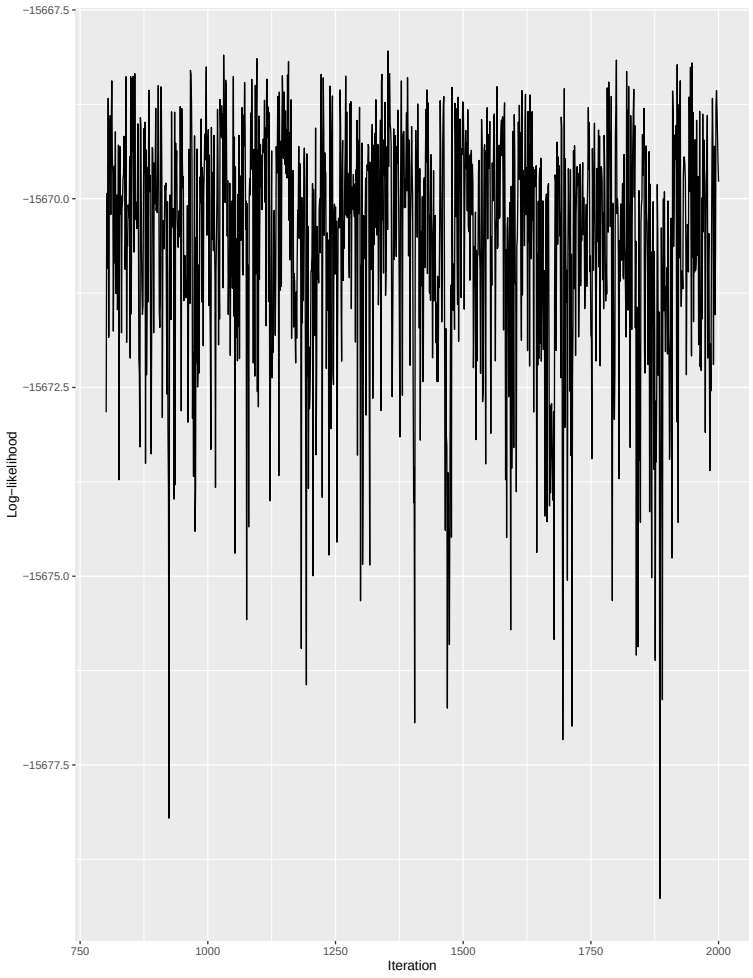

Figure 11: Trace of the log-likelihood, after discarding the burnin.

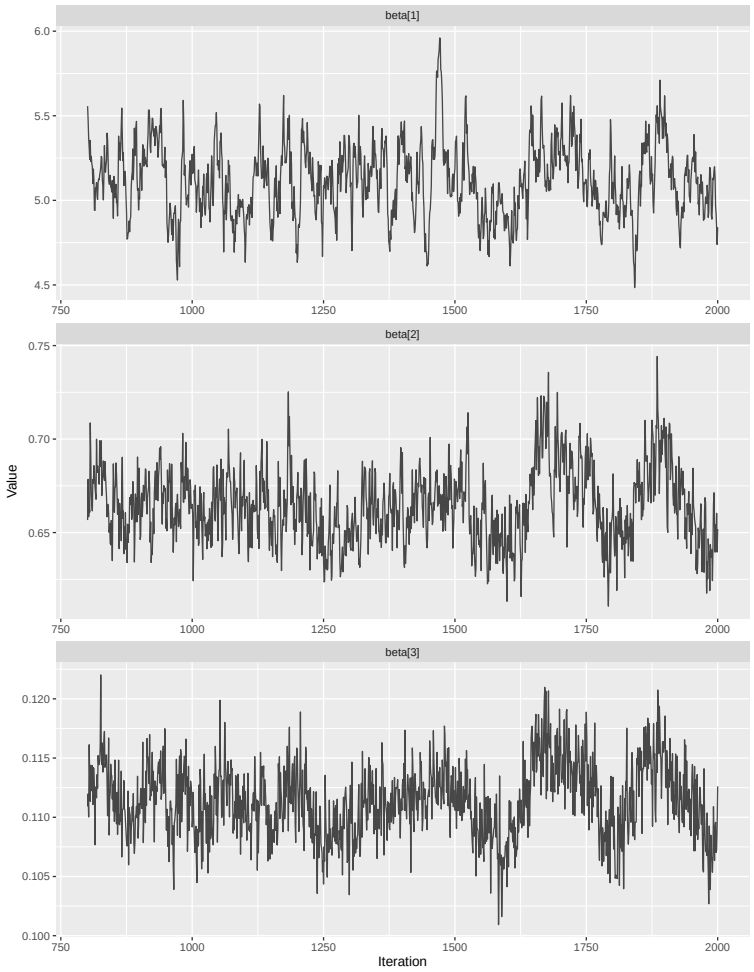

Figure 12: Trace of the individual inferred betas, after discarding the burnin.

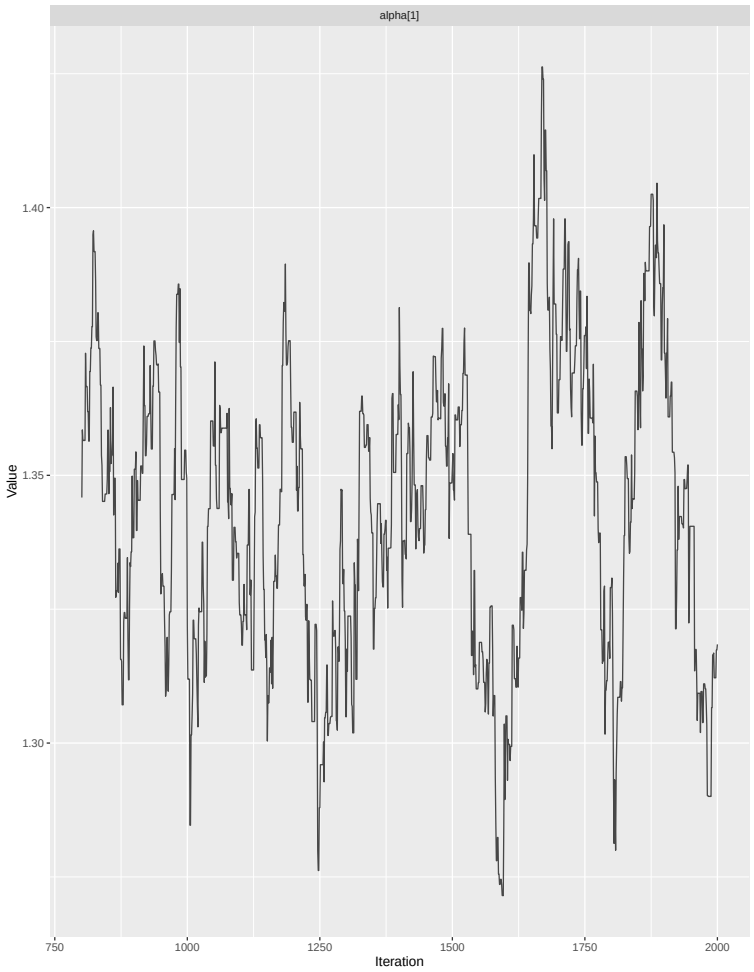

Figure 13: Trace of the inferred alpha, after discarding the burnin.

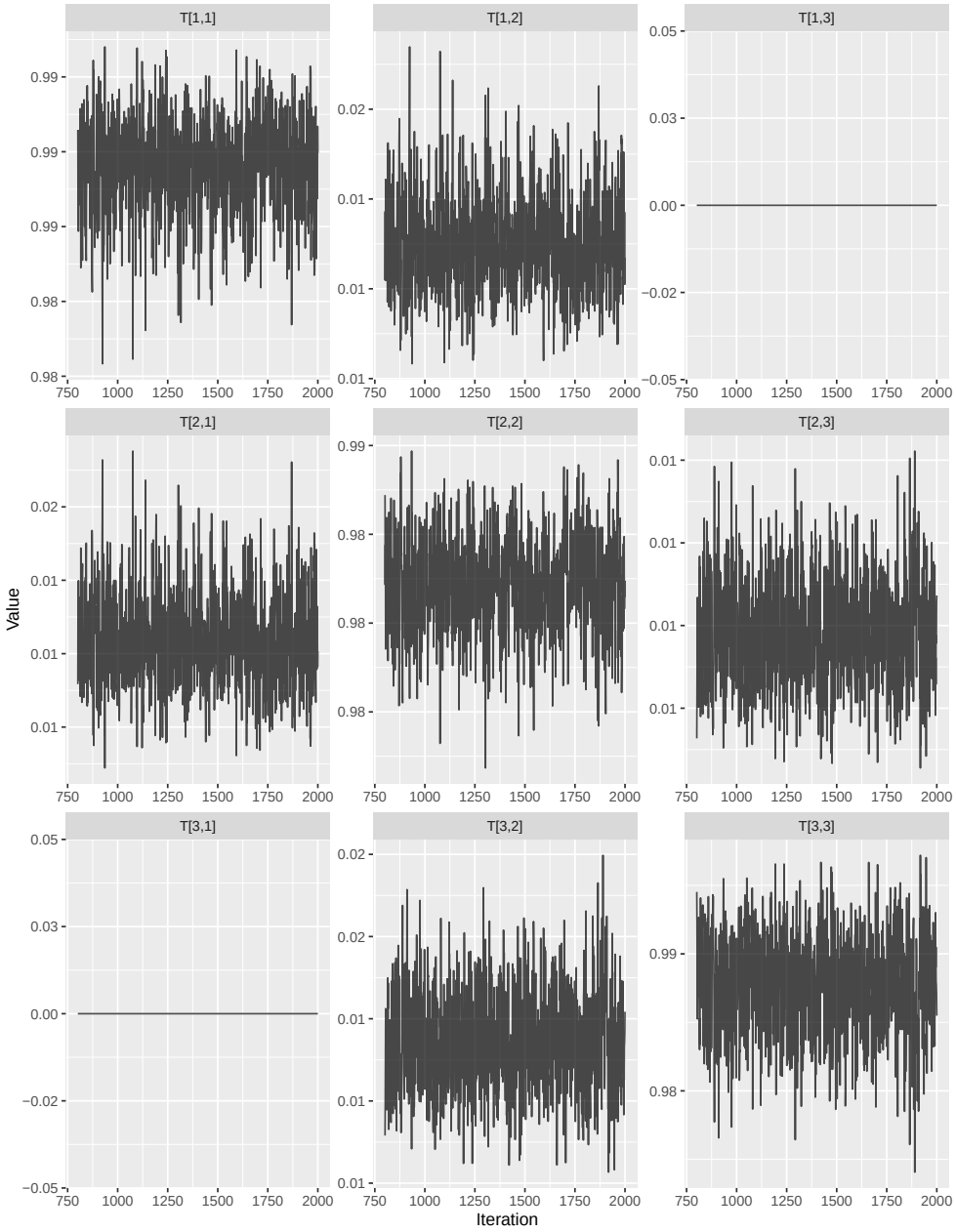

Figure 14: Traces of the inferred entries of the transition rate matrix, after discarding the burnin.

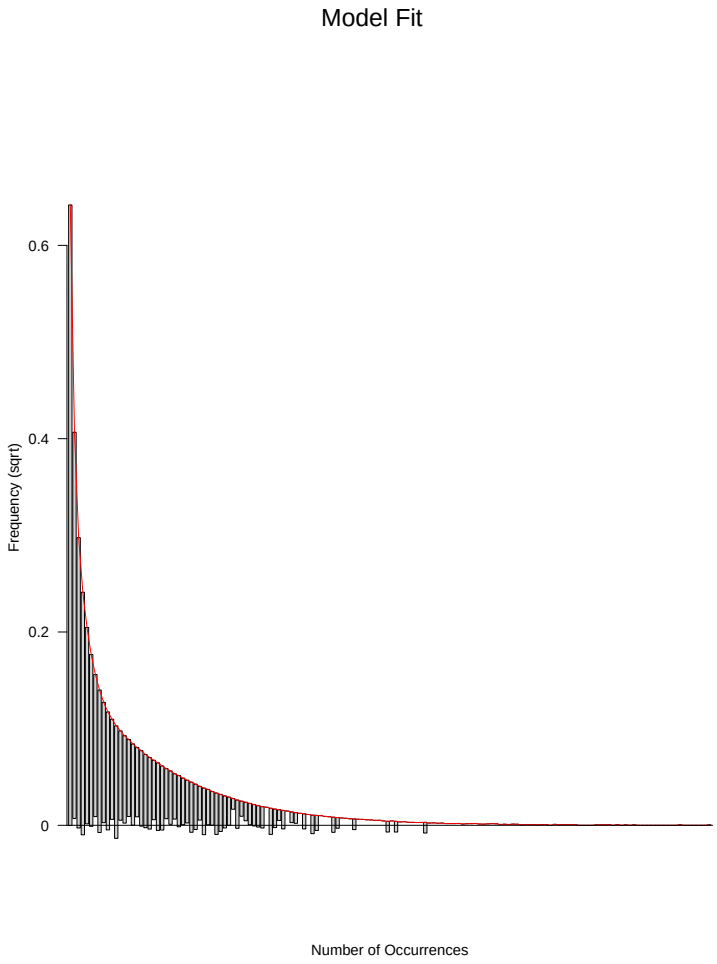

Figure 15: Rootogram of the observed data (vertical bars) fitted to the inferred mixture of poisson distributions (smoothed red line).

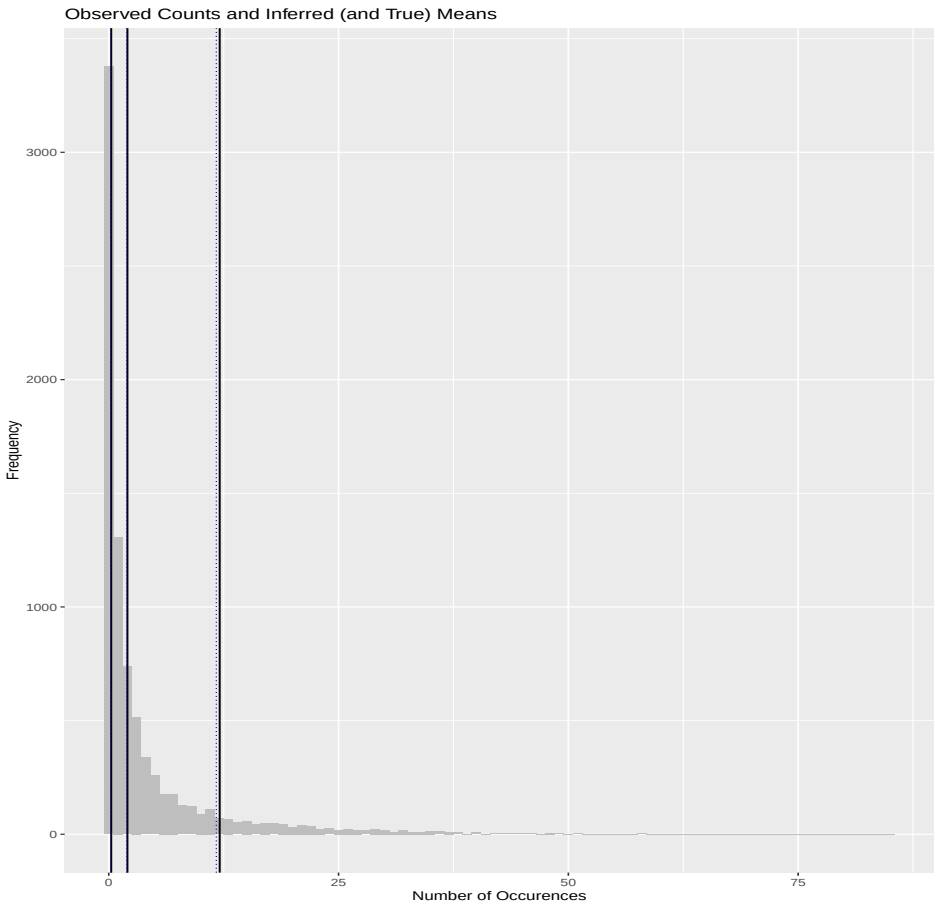

Figure 16: Histogram of the observed data with the true means in the solid vertical lines, and the inferred means as dotted vertical means.

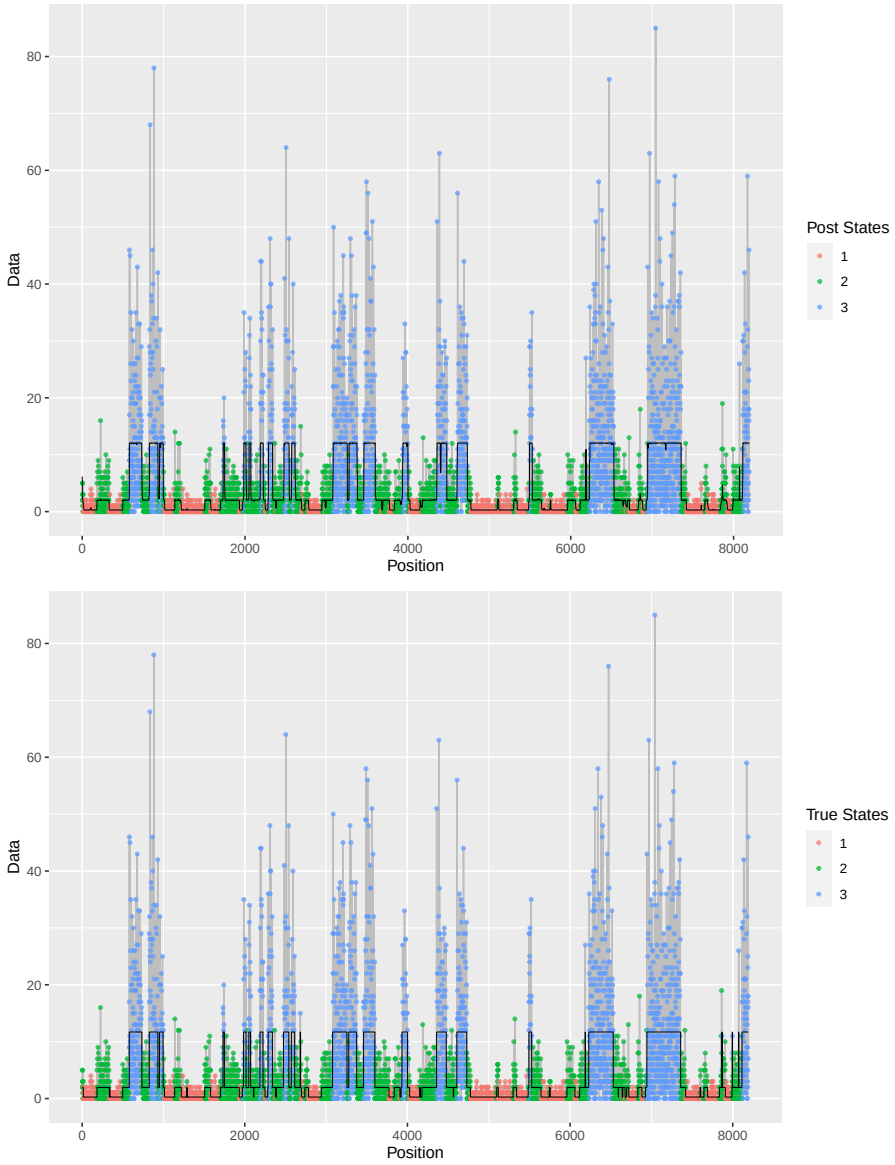

Figure 17: These plots show the data points/emissions for each segment along the sequence coloured by the hidden states. In the upper plot, these are the true hidden mean; in the lower plot, these are the inferred hidden states. The black lines trace the inferred posterior mean across the sequence. Recall 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.