Fast sampling and model selection for Bayesian mixture models: Supplementary Information

M. E. J. Newman

Center for the Study of Complex Systems, University of Michigan, Ann Arbor, 48109, Michigan, USA.

Corresponding author(s). E-mail(s): mejn@umich.edu;

S1 Proof of correctness for the Monte Carlo algorithm

In this section we prove that the distribution of samples drawn by the algorithm of Section 3 converges to the integrated posterior distribution of the corresponding mixture model.

S1.1 Reformulation of the algorithm

To prove convergence, we first give an alternate formulation of the algorithm, which would be less efficient in practice but for which the proof is simpler. After proving the correctness of this algorithm we then demonstrate that the algorithm of Section 3—the one we actually use—is equivalent to this alternate formulation.

A step of the alternate algorithm is as follows.

1. Choose a component r uniformly at random, then choose an observation i uniformly at random from that component.
2. Remove i from component r.
3. If i was the only member of component r, delete component r, relabel component k to be the new component r, and decrease k by 1. (If $r = k$ then no relabeling is necessary; we simply delete the component.)
4. Assemble a set of $2k + 1$ candidate states, each of which has a weight associated with it, as follows.
   (a) Of the $2k + 1$ states, k of them are states in which i is placed in one of the k current components s. The weights associated with each of these candidate

states are

$$w_\mu = \frac{P(x|k, z_{-i}, z_i = s)}{P(x|k, z_{-i})}, \tag{S1}$$

where z_{-i} denotes all $N-1$ component assignments except for z_i.

(b) The remaining $k+1$ candidate states are ones that make i the sole member of a new component $k+1$, swap the labels of component $k+1$ and another component $s = 1 \ldots k+1$, then increase k by 1. (If $s = k+1$ no swapping of labels is necessary; we simply create the new component.) The weight associated with each of these states is

$$w_\mu = \frac{k^2}{(k+1)(N-k)} \frac{P(k+1)P(x|k+1, z_{-i}, z_i = s)}{P(k)P(x|k, z_{-i})}, \tag{S2}$$

where $P(k)$ is the prior on k as previously, k is the number of components before the new component is added, and the swapping of the labels is already incorporated into the values of z_{-i}.

5. Once the set of candidate states is constructed, choose one state μ in proportion to its weight, i.e., with probability

$$p_\mu = \frac{w_\mu}{\sum_\nu w_\nu}, \tag{S3}$$

and update the system to that new state.

S1.2 Ergodicity

Proof of convergence for any Markov chain Monte Carlo algorithm requires that the algorithm satisfy two conditions: ergodicity and detailed balance (Newman and Barkema, 1999). The requirement of ergodicity states that the algorithm must be able to reach any state of the system from any other in a finite number of Monte Carlo steps. In our algorithm, the states are pairs k, z of number of components plus component assignment for all observations. For these states ergodicity is trivially satisfied by the algorithm above, since our basic Monte Carlo moves can move any observation to a different component or create or delete components, and a finite series of such moves can reach a state with any value of k and component assignment z.

S1.3 Detailed balance

The more demanding step is the proof of detailed balance. Detailed balance is the requirement that for any pair of states μ, ν the average rate of transitions between them is the same in either direction in equilibrium. If $P(\mu)$ is the equilibrium probability of being in state μ and $P(\mu \to \nu)$ is the probability of making a transition from μ to ν, then the detailed balance condition can be written in the form

$$P(\mu)P(\mu \to \nu) = P(\nu)P(\nu \to \mu), \tag{S4}$$

or equivalently

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{P(\nu)}{P(\mu)}. \tag{S5}$$

To prove that our algorithm satisfies detailed balance we must consider several cases. First, consider moves that do not change the number of components. There are two types of such moves. The first are trivial moves in which we remove the last member i of a component r and delete the component, but then immediately create a new component s, making i its sole member. Detailed balance is simple in this case because the reverse of such a move is the same move again, and hence the probabilities $P(\mu \to \nu)$ and $P(\nu \to \mu)$ are equal, which trivially satisfies detailed balance.

More complicated are the steps where we move an observation from one component to another without creating or deleting any components. The probability $P(\mu \to \nu)$ for the forward move in this case is equal to the probability $1/k$ of choosing a particular component r, times the probability $1/n_r$ of picking a particular observation i from that component, times the probability $p_\nu = w_\nu / \sum_\mu w_\mu$ of choosing the target state ν:

$$P(\mu \to \nu) = \frac{1}{k} \times \frac{1}{n_r} \times \frac{w_\nu}{\sum_\mu w_\mu}. \tag{S6}$$

For the backward move the corresponding probability is

$$P(\nu \to \mu) = \frac{1}{k} \times \frac{1}{n'_s} \times \frac{w_\mu}{\sum_\mu w_\mu}, \tag{S7}$$

where the primed variable n'_s denotes the value in state ν. The sums in the denominators of (S6) and (S7) are over the same set of states and so have the same value, and hence, using Eq. (S1) for the weights, the ratio of probabilities is

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{n'_s}{n_r} \frac{w_\nu}{w_\mu} = \frac{n_s + 1}{n_r} \frac{P(x|k, z_{-i}, z_i = s)}{P(x|k, z_{-i}, z_i = r)}. \tag{S8}$$

where the remaining factors have canceled and we have made use of $n'_s = n_s + 1$.

To prove detailed balance we need to show that this ratio is equal to the ratio of posterior probabilities $P(\nu)/P(\mu)$ as in Eq. (S5). Using Eq. (12), we have

$$P(k, z|x) = \frac{P(k)P(x|k, z)}{P(x)} \binom{N-1}{k-1}^{-1} \frac{\prod_r n_r!}{N!}. \tag{S9}$$

Taking the ratio of probabilities for the states before and after the move, many factors cancel and we get

$$\frac{P(\nu)}{P(\mu)} = \frac{n'_r! n'_s!}{n_r! n_s!} \frac{P(k)P(x|k, z_{-i}, z_i = s)/P(x)}{P(k)P(x|k, z_{-i}, z_i = r)/P(x)} = \frac{n_s + 1}{n_r} \frac{P(x|k, z_{-i}, z_i = s)}{P(x|k, z_{-i}, z_i = r)}, \tag{S10}$$

where we have used $n'_r = n_r - 1$ and $n'_s = n_s + 1$.

Equation (S10) is indeed equal to Eq. (S8) and hence detailed balance is established. The only exception is for the special case where an observation is "moved" to the same component it is already in, so that no change of state occurs. However, it is easy to demonstrate that, with the choice of weights in Eq. (S1), both sides of Eq. (S5) are equal to 1 in this case, so again detailed balance is satisfied.

The last class of moves we need to consider are those that do change the number of components, either deleting a component or adding a new one. For the purposes of our proof, and without loss of generality, let us consider the forward move $\mu \to \nu$ to be the one that deletes a component and the reverse move to be the one that adds a component. Thus, the forward move removes the last observation i from component r and, since the component is now empty, deletes the component and replaces it with the component that was previously labeled k, then decreases the value of k by one. Then i is placed in one of the other current components s. The probability of this move is equal to the probability $1/k$ of choosing component r times the probability $p_\nu = w_\nu / \sum_\mu w_\mu$ of choosing to place i in component s:

$$P(\mu \to \nu) = \frac{1}{k} \times \frac{w_\nu}{\sum_\mu w_\mu}, \tag{S11}$$

where w_ν is as in Eq. (S1) and k denotes the number of components before r is deleted.

The reverse move removes observation i from component s and makes it the sole member of a new component r. The probability of this move is equal to the probability $1/k'$ of choosing s from the k' possibilities, times the probability $1/n'_s$ of picking observation i, times the probability p_μ of choosing the move that labels the new component as component r, where once again the primed variables denote values in state ν. Thus,

$$P(\nu \to \mu) = \frac{1}{k'} \times \frac{1}{n'_s} \times \frac{w_\mu}{\sum_\mu w_\mu}. \tag{S12}$$

Again the sums in the denominators of (S11) and (S12) are equal and, using Eqs. (S1) and (S2), the ratio of probabilities is

$$\begin{aligned}
\frac{P(\mu \to \nu)}{P(\nu \to \mu)} &= \frac{k'}{k} n'_s \frac{w_\nu}{w_\mu} \\
&= \frac{k-1}{k}(n_s+1)\frac{P(x|k-1, z_{-i}, z_i = s)}{P(x|k-1, z_{-i})} \frac{(N-k+1)k}{(k-1)^2} \frac{P(k-1)P(x|k-1, z_{-i})}{P(k)P(x|k, z_{-i}, z_i = r)} \\
&= \frac{N-k+1}{k-1}(n_s+1)\frac{P(k-1)P(x|k-1, z_{-i}, z_i = s)}{P(k)P(x|k, z_{-i}, z_i = r)}.
\end{aligned} \tag{S13}$$

Meanwhile, using Eq. (S9), the ratio of posterior probabilities for the two states, is

$$\begin{aligned}
\frac{P(\nu)}{P(\mu)} &= \frac{P(k-1)P(x|k-1, z_{-i}, z_i = 2)/P(x)}{P(k)P(x|k, z_{-i}, z_i = r)/P(x)} \frac{(k-2)!(N-k+1)!/(N-1)!}{(k-1)!(N-k)!/(N-1)!} \frac{(n_s+1)!}{n_s!} \\
&= \frac{N-k+1}{k-1}(n_s+1)\frac{P(k-1)P(x|k-1, z_{-i}, z_i = s)}{P(k)P(x|k, z_{-i}, z_i = r)},
\end{aligned} \tag{S14}$$

which is equal to (S13) and hence detailed balance is again established. This completes the proof of correctness for the algorithm of this section.

Finally, we observe that the moves that create new components with labels $s = 1 \dots k+1$ are all equivalent up to a label permutation and hence, if we don't care about such permutations, we can lump them all together and represent them with a single move that creates a new component $k+1$ and carries $k+1$ times the weight, Eq. (S2), of each individual move, which gives

$$w_\mu = \frac{k^2}{N-k} \frac{P(k+1)P(x|k+1, z_{-i}, z_i = k+1)}{P(k)P(x|k, z_{-i})}, \tag{S15}$$

as in Eq. (15). This is the algorithm described in Section 3 and the one we use in our calculations. Lumping states together like this means that when the algorithm performs a move followed by its reverse move, we may end up not in the state we started with but in an equivalent state with permuted labels. This has no effect on the division of the observations into components or on our estimate of the number of components, but it saves us some time and complexity in the algorithm.

S1.4 Implementation

A number of points about implementation of the algorithm are worth mentioning. First, it is inefficient to implement it in terms of the component membership variables z_i. A better approach is to maintain instead a list in a simple array of the observations in each component, along with a record of the number n_r of members of the component. This allows one to choose a random member of a component quickly, as required by the algorithm, and a member can be efficiently added to a component by putting it in the first free element of the array. A member can be removed by overwriting it with the last element.

Some Monte Carlo steps also require us to renumber entire components. Rather than renumbering each member of a component separately, which would be slow, we instead simply swap pointers to the memory locations containing the membership lists, which can be done in O(1) time.

Some computational effort can also be saved by delaying the removal of a selected observation from its component. It is straightforward to calculate the weights in Eqs. (14) and (15) as if the observation *had* been removed, which allows us to decide which move to perform before making any updates. With these precautions, the running time for a single Monte Carlo step of the algorithm can be reduced to $O(k)$, which is optimal since k weights have to be computed for every step.

S2 Additional examples

In this section we give additional example applications of the LCA algorithm from Section 4.3 to real-world data sets, including one data set significantly larger than any other we consider.

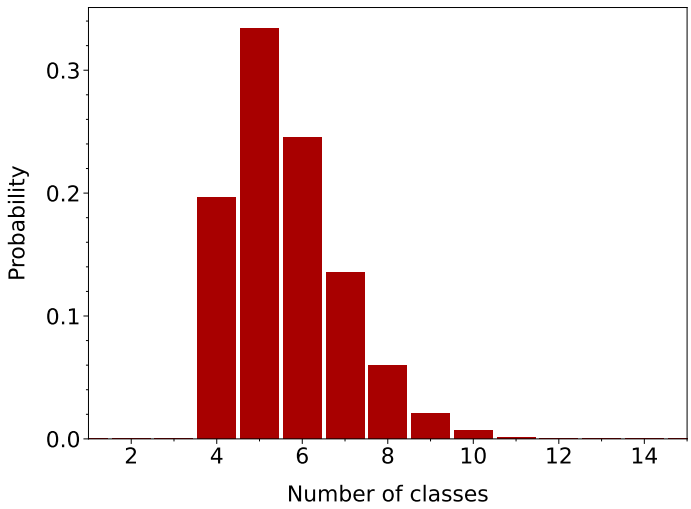

Fig. S1 Posterior distribution of the number of classes in the teenage behavior data set.

S2.1 Problem behaviors among teenagers

This example is an application to a large behavioral data set and illustrates a case in which our method gives different answers from previous approaches. We examine a data set presented by Li et al (2018), which describes survey results on the presence or absence of six problem behaviors in 6504 US teenagers. A previous analysis of a subset of these data by Collins and Lanza (2010) found four latent classes. Figure S1 shows the posterior distribution of the number of classes returned by our Monte Carlo method and, as we can see, four classes is a possible fit in this case, but the MAP estimate is five, and six is also more likely than four. We find a longer run is necessary with this data set than for others we consider, in order to achieve consistent results—we ran for 250 000 Monte Carlo sweeps, plus 25 000 for burn-in. Nonetheless the calculation is fast: Li et al. report that their Gibbs sampling calculation took about half an hour; ours takes less than 90 seconds on roughly comparable hardware.

S2.2 Online dating

In this example we demonstrate an application of the method to a much larger data set, which comes from the online dating service OKCupid. Upon creating accounts on this service, users are asked a set of questions about themselves, which the service uses to match them with potential partners. Our data set consists of answers from about 60 000 users to 14 of these questions, which ask about gender, sexual orientation, relationship status, income, offspring, body type, diet, drinking and smoking habits, drug use, zodiac sign, religion, and whether they like cats and dogs.

The analysis of this data set takes considerably longer than any of the previous ones, since one sweep now corresponds to 60 000 Monte Carlo steps and hence takes longer to complete. 25 000 sweeps, for example, takes about nine minutes on the author's laptop. Figure S2 (main panel) shows the posterior distribution of the number of classes and, as we can see, this data set has a much larger number of classes

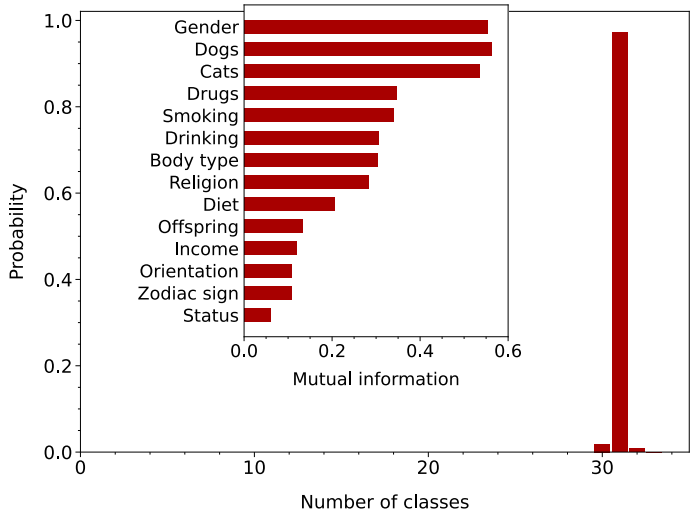

Fig. S2 Main figure: Distribution of inferred number of classes in the online dating data set. Inset: Mutual information between data and class assignments for each of the 14 questions.

than any of our others, with the distribution strongly peaked around a MAP estimate of $k = 31$ classes. Inset in the figure we show the mutual information scores for the 14 questions, calculated using the method of Section 4.5.3. As before, these scores indicate which observations are most informative about the component assignment. The most informative observations in this case are (perhaps unsurprisingly) gender and (more surprisingly) whether the respondents like cats and dogs. Apparently the world really is divided between cat people and dog people. High mutual information scores also go to the questions on smoking, drinking, and drugs (which may well be correlated with one another) and to body type and religion. At the bottom of the list are zodiac sign, which is perhaps not surprising, but also status (single, married, etc.) and sexual orientation. The latter seem more surprising, but we should bear in mind that their low mutual information scores merely indicate that they are not correlated with people's answers to the other questions. They are saying, for example, that being straight or gay doesn't affect whether you drink or like dogs.

S3 Algorithm for the model with general η

The algorithm given in Section 3, as well as our example calculations, all assume a Dirichlet-categorical prior over component assignments with concentration parameter $\eta = 1$. The methods of this paper can, however, be extended to general values of η, for which the prior takes the form given in Eq. (8):

$$P(z|k,\eta) = \frac{1}{N!}(N-k)\,\mathrm{B}(N-k,k\eta)\prod_{r=1}^{k} n_r \frac{\Gamma(n_r+\eta-1)}{\Gamma(\eta)}, \tag{S16}$$

where $\mathrm{B}(x,y)$ is Euler's beta function. The generalization makes use of a Bortz-Kalos-Lebowitz style continuous-time Monte Carlo dynamics (Bortz et al, 1975; Newman and Barkema, 1999) in which we keep track of a real-valued time variable t that measures elapsed time in arbitrary units. For general η, a step of the algorithm is as follows:

1. Define a set of weights u_r for $r = 1 \ldots k$ by

$$u_r = \begin{cases} 1 & \text{if } n_r = 1, \\ (n_r - 1)/(n_r + \eta - 2) & \text{if } n_r > 1, \end{cases} \tag{S17}$$

   then draw a component r at random from the categorical distribution with probabilities $u_r / \sum_s u_s$. Simultaneously, increase the time variable according to

$$t \to t + \frac{k}{\sum_s u_s}. \tag{S18}$$

2. Choose an observation i uniformly at random from component r.
3. Remove i from component r.
4. If i was the only member of component r, delete component r, relabel component k to be the new component r, and decrease k by 1. (If $r = k$ then no relabeling is necessary.)
5. Assemble a set of $k+1$ candidate states $\mu = (k, z)$, each of which has a weight w_μ associated with it, as follows.
   (a) Of the $k+1$ states, k of them are states in which i is placed in one of the k current components s. Each of these candidate states has an associated weight

$$w_\mu = \frac{P(x|k, z_{-i}, z_i = s)}{P(x|k, z_{-i})}, \tag{S19}$$

   where as previously z_{-i} denotes the set of all component assignments except for z_i.

   (b) The last candidate state is one that makes i the sole member of a new component $k+1$ and increases k by 1. The weight associated with this state is

$$w_\mu = \frac{k(N-k-1)\,\mathrm{B}(N-k-1,(k+1)\eta)}{(N-k)\,\mathrm{B}(N-k,k\eta)} \frac{P(k+1)P(x|k+1, z_{-i}, z_i = k+1)}{P(k)P(x|k, z_{-i})}, \tag{S20}$$

   where k is the number of components before the new component is added.
6. Once the set of target states is assembled, choose one of them μ with probability p_μ, Eq. (S3), and update the system to that new state.

Sample states are drawn in the normal manner at uniform intervals, but using time as measured by the time variable t, rather than time in Monte Carlo steps. For the conventional choice $\eta = 1$, this algorithm reduces to the algorithm of the main paper. (We leave the demonstration as an exercise for the interested reader.)

Proof of the correctness of the algorithm follows similar lines to that of Section S1.3, but allowing for the effect of the continuous time variable. For moves that do not

change the number of components, merely moving an observation from one component to another, the probability for the forward move from component r to component s is equal to the probability of choosing component r, times the probability $1/n_r$ of picking the particular observation i, times the probability $p_\nu = w_\nu / \sum_\mu w_\mu$ of choosing the candidate state ν, which gives

$$P(\mu \to \nu) = \frac{u_r}{\sum_r u_r} \times \frac{1}{n_r} \times \frac{w_\nu}{\sum_\mu w_\mu}. \tag{S21}$$

For the backward move the probability is

$$P(\nu \to \mu) = \frac{v_s}{\sum_s v_s} \times \frac{1}{n'_s} \times \frac{w_\mu}{\sum_\mu w_\mu}, \tag{S22}$$

where v_s are the weights of Eq. (S17) for the backward move. Bearing in mind that n_r and n'_s must both be greater than 1 if the number of components does not change, and making use of Eq. (S17) for u_r and v_s, the ratio of forward and backward probabilities is then

$$\begin{aligned} \frac{P(\mu \to \nu)}{P(\nu \to \mu)} &= \frac{u_r / \sum_r u_r}{v_s / \sum_s v_s} \frac{n'_s}{n_r} \frac{w_\nu}{w_\mu} \\ &= \frac{(n_r - 1)/n_r}{n_s/(n_s+1)} \left(\frac{n_s + \eta - 1}{n_r + \eta - 2} \right) \frac{P(x|k, z_{-i}, z_i = s)}{P(x|k, z_{-i}, z_i = r)} \frac{\sum_s v_s}{\sum_r u_r}. \end{aligned} \tag{S23}$$

Now we compare this to the ratio of the desired equilibrium probabilities of the corresponding states. We have

$$P(k, z|x, \eta) = \frac{P(x|k, z) P(z|k, \eta) P(k)}{P(x)}, \tag{S24}$$

and hence, using (S16), we have

$$\begin{aligned} \frac{P(\nu)}{P(\mu)} &= \frac{n'_r \Gamma(n'_r + \eta - 1) n'_s \Gamma(n'_s + \eta - 1) P(x|k, z_{-i}, z_i = s) P(k)/P(x)}{n_r \Gamma(n_r + \eta - 1) n_s \Gamma(n_s + \eta - 1) P(x|k, z_{-i}, z_i = r) P(k)/P(x)} \\ &= \frac{(n_r - 1)/n_r}{n_s/(n_s+1)} \left(\frac{n_s + \eta - 1}{n_r + \eta - 2} \right) \frac{P(x|k, z_{-i}, z_i = s)}{P(x|k, z_{-i}, z_i = r)}. \end{aligned} \tag{S25}$$

Substituting this expression into Eq. (S23) we have

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{P(\nu)\,(1/k) \sum_s v_s}{P(\mu)\,(1/k) \sum_r u_r}, \tag{S26}$$

which is similar to the standard detailed balance formula, Eq. (S5), except for the factors of $(1/k)\sum_r u_r$ and $(1/k)\sum_s v_s$. This means that each state μ will appear not with its normal probability $P(\mu)$ but with modified probability $P(\mu)(1/k)\sum_r u_r$. However, the probability of *sampling* from state μ is multiplied by the amount of time

for which the system remains in that state, which is $k/\sum_r u_r$, following Eq. (S18). So the probability of sampling is precisely proportional to $P(\mu)$, as required.

For Monte Carlo steps that change the number of components, we can once more, without loss of generality, consider the forward move $\mu \to \nu$ to be the one that removes the last observation i from a component r and then deletes the empty component, and the reverse move to be the one that creates a new component. The probability of the forward move is equal to the probability of choosing component r times the probability $p_\nu = w_\nu / \sum_\mu w_\mu$ of placing observation i in a particular new component s:

$$P(\mu \to \nu) = \frac{u_r}{\sum_r u_r} \times \frac{w_\nu}{\sum_\mu w_\mu}. \tag{S27}$$

The backward move removes observation i from component s and makes it the sole member of a new component r. The probability for this move is equal to the probability of choosing component s, times the probability $1/n'_s$ of picking observation i, times the probability p_μ of choosing the move that creates the new component. As with the proof in Section S1, we will initially assume separate weights for creating new components with each possible component label $1 \ldots k'+1$ and values

$$\begin{aligned} w_\mu &= \frac{k'(N-k'-1)\,\mathrm{B}(N-k'-1,(k'+1)\eta)}{(k'+1)(N-k')\,\mathrm{B}(N-k',k'\eta)} \frac{P(k'+1)P(x|k'+1,z_{-i},z_i=r)}{P(k')P(x|k',z_{-i})} \\ &= \frac{(k-1)(N-k)\,\mathrm{B}(N-k,k\eta)}{k(N-k+1)\,\mathrm{B}(N-k+1,(k-1)\eta)} \frac{P(k)P(x|k,z_{-i},z_i=r)}{P(k-1)P(x|k-1,z_{-i})}. \end{aligned} \tag{S28}$$

Then we have

$$P(\nu \to \mu) = \frac{v_s}{\sum_s v_s} \times \frac{1}{n'_s} \times \frac{w_\mu}{\sum_\mu w_\mu}, \tag{S29}$$

and the ratio of probabilities is

$$\begin{aligned} \frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{u_r/\sum_r u_r}{v_s/\sum_s v_s} n'_s \frac{w_\nu}{w_\mu} &= \frac{n_s+\eta-1}{n_s} \frac{\sum_s v_s}{\sum_r v_r} (n_s+1) \\ &\times \frac{k(N-k+1)\,\mathrm{B}(N-k+1,(k-1)\eta)}{(k-1)(N-k)\,\mathrm{B}(N-k,k\eta)} \frac{P(k-1)P(x|k-1,z_{-i},z_i=s)}{P(k)P(x|k,z_{-i},z_i=r)}. \end{aligned} \tag{S30}$$

Meanwhile, the ratio of equilibrium probabilities for states μ and ν is

$$\begin{aligned} \frac{P(\nu)}{P(\mu)} &= \frac{(N-k+1)\,\mathrm{B}(N-k+1,(k-1)\eta)}{(N-k)\,\mathrm{B}(N-k,k\eta)} \frac{(n_s+1)\Gamma(n_s+\eta)}{\Gamma(\eta)} \frac{\Gamma(\eta)^2}{\Gamma(\eta)n_s\Gamma(n_s+\eta-1)} \\ &\times \frac{P(x|k-1,z_{-i},z_i=s)P(k-1)/P(x)}{P(x|k,z_{-i},z_i=r)P(k)/P(x)} \\ &= (n_s+\eta-1)\frac{n_s+1}{n_s} \frac{(N-k+1)\,\mathrm{B}(N-k+1,(k-1)\eta)}{(N-k)\,\mathrm{B}(N-k,k\eta)} \\ &\times \frac{P(k-1)P(x|k-1,z_{-i},z_i=s)}{P(k)P(x|k,z_{-i},z_i=r)}. \end{aligned} \tag{S31}$$

Comparing with Eq. (S30), we find that

$$\frac{P(\nu)}{P(\mu)} = \frac{P(\nu)[1/(k-1)]\sum_s v_s}{P(\mu)(1/k)\sum_r u_r}. \tag{S32}$$

Hence, once again, we have a modified detailed balance condition that implies that each state μ will appear with probability $P(\mu)(1/k)\sum_r u_r$. But the probability of *sampling* each state is multiplied by the length of time for which the system remains in that state, which is $k/\sum_r u_r$ following Eq. (S18), and hence the probability of sampling is exactly proportional to $P(\mu)$, as required.

The final step of the proof, as in Section S1, is to combine the $k+1$ moves that create new components numbered $1 \dots k+1$ into a single move that creates a component numbered $k+1$, with $k+1$ times the probability. The weight for this move is given by Eq. (S28) times $k'+1$, which gives the expression in Eq. (S20). As previously, this can result in a permutation of the labels of the components, but it has no effect on the actual partition of the observations, since the component labels are arbitrary.

This completes the proof of correctness for the generalized algorithm.

Implementation of the algorithm is straightforward and the slightly higher level of complexity compared with the case for $\eta = 1$ does not impact performance significantly. In particular, note that one can calculate in advance tables of $n/(n+\eta-1)$ and $\mathrm{B}(N-k, k\eta)$ for $n, k = 1 \dots N$ and then use them to evaluate Eqs. (S17) and (S19) with little performance penalty.

S4 Data sets

In this section we describe the data sets used in our examples.

S4.1 Epileptic seizures

These data come from a clinical trial of medications for pediatric epilepsy reported by Wang et al (1996). The data describe the experiences of a single epileptic child participant in the trial over a 140-day period. The participant was observed under baseline conditions for 28 days, then received monthly infusions of intravenous gammaglobulin, as a potential treatment, for the remainder of the observation period, and the data set consists of the number of seizure episodes experienced on each day of the trial, as recorded in a diary by the child's parent. The complete data set is included in the paper by Wang et al.

S4.2 Candy dispenser

This data set comes from an automated candy dispenser in the reception area outside the author's office, which is stocked with M&Ms brand candy. The dispenser uses a motorized corkscrew mechanism to dispense a handful of M&Ms upon the push of a button. The data set records the number of candies dispensed on 857 pushes of the button and was gathered by Max Jerdee, Conrad Kosowski, Gabriela Fernandes Martins,

Table 1 The 32 questions selected from the CBS/New York Times poll for the data set used in this paper, listed in the order in which they were asked on the survey.

Item	Question	Responses
`sex`	Respondent's sex	2
`cenr`	Census region	4
`q1`	Obama job approval	3
`q2`	Right direction/wrong track	3
`q6`	Congress job approval	3
`q13`	Rate national economy	5
`reg`	Registered voter	3
`q66`	Spending cuts vs. jobs	4
`q67`	Payroll tax cut	3
`q69`	Spending on infrastructure	3
`q70`	Small business tax cut	3
`q76`	Tax on wealthy	3
`q78`	Economic liberal/conservative	6
`q79`	Social liberal/conservative	6
`q80`	Social security/Medicare	3
`q84`	Repeal health care law	4
`q86`	Death penalty	3
`q88`	Global warming	6
`q89`	Same-sex marriage	4
`q90`	Abortion	4
`q92`	Illegal immigration	4
`prty`	Party ID	4
`q106`	Employment status	5
`vt08`	Voted for in 2008	6
`evan`	Evangelical	3
`reli`	Religion	7
`marr`	Marital status	6
`agea`	Age group	5
`educ`	Education	6
`hisp`	Hispanic	3
`race`	Race	5
`inca`	Income group	6

and Chethan Prakash. The measurements were near-consecutive, but four measurements had to be discarded because of missing data, leaving a total of 853 for analysis. A copy of the data set is available for download at `https://umich.edu/~mejn/mixture`.

S4.3 CBS/New York Times poll

These data come from an opinion poll, fielded jointly by the CBS television network and the New York Times newspaper, of 1566 members of the American public during September 2011, and focusing on a range of topics of then-current interest. The data set we examine contains responses for all participants but only a 32-question subset of the questions asked, as listed in Table 1. The data were provided by the Inter-university Consortium for Political and Social Research (ICPSR) and are freely available from the ICPSR web site, file number 34458, at `https://doi.org/10.3886/ICPSR34458.v1`.

Table 2 Summary of the 12 variables used in the 50 states data set. "College education" measures the fraction of the population over the age of 25 with a bachelor's degree. "Football" refers to whether the state is home to a team in the US National Football League. "Gun laws" distinguishes constitutional carry, concealed carry plus unlicensed open carry, concealed carry plus licensed open carry, and concealed carry only. "Medicaid" refers to adoption of the Medicaid expansion under the Affordable Care Act of 2010, at the time of writing of this paper. "Same-sex marriage" refers to state statutes and constitutions that contain, or do not contain, wording forbidding same-sex marriage. Actual same-sex marriage has been legal nationwide in the US since 2015 under federal law, which preempts state laws.

Variable	Values
Abortion	Legal, Limited, Illegal
Cannabis	Legal, Limited, Illegal
Census region	Northeast, Midwest, South, West
College education	Fraction with bachelor's: below 30%, 30–35%, 35–40%, above 40%
Death penalty	Yes, not enforced, no
Football	Has an NFL team, does not have a team
Gun laws	Yes, unlicensed open, licensed open, concealed carry only
Medicaid	Expanded under the ACA, not expanded
Sales tax	Below 1.75%, 1.75–7.07%, above 7.07%
Same-sex marriage	Banned by statute, not banned
Temperature	Average temperature: below 5°C, 5–15°C, above 15°C
Vote 2024	Harris, Trump

S4.4 Alzheimer's diagnostic survey

This data set, which comes from Walsh (2006), records the presence or absence of six symptoms of potential Alzheimer's disease in 240 patients at the Mercer's Institute national memory clinic in Dublin, Ireland, as reported by the patients' primary caregivers on the occasion of the patients' first visit to the clinic. The patients were preselected as having suspected Alzheimer's disease or other age-related cognitive decline, but only mild symptoms, so that a positive diagnosis was not unlikely but also not guaranteed. The symptoms identified were: activity disturbance, affective disorder, aggression, agitation, diurnal rhythm disturbance, and hallucinations. The survey, a standard diagnostic instrument called the Behave-AD questionnaire, returns information on both the presence and the severity of each symptom, but Walsh converted the data to dichotomous presence/absence variables and this is the version we analyze. The data set is included in full in a table within the paper by Walsh.

S4.5 50 states data set

This data set, which was assembled by the current author from online sources for the purposes of this study, records 12 traits for each of the 50 United States (but excluding the District of Columbia, Puerto Rico, and other non-state territories). The 12 traits are listed in Table 2 and further details are given in the table caption. The full data set is available for download from `https://umich.edu/~mejn/mixture`.

S4.6 Problem behaviors among teenagers

This data set comes from Li et al (2018) and describes self-reported incidence of problem behaviors by 6504 teenagers in the United States during the 1990s. The

Table 3 Summary of the 14 variables in the online dating data set. Missing data for each question is also coded as an additional possible answer.

Variable	Values
Gender	Male, female
Orientation	Straight, gay, bisexual
Status	Single, available, seeing someone, married
Income	Less than $20k, over $20k, $30k, $40k, $50k, $60k, $70k, $80k, $100k, $150k, $250k, $500k, $1m
Offspring	Has children, doesn't have children, wants children, might want children, doesn't want children
Body type	Skinny, thin, average, fit, athletic, a little extra, curvy, full figured, jacked, used up, overweight, rather not say
Diet	Anything, vegetarian, vegan, kosher, halal, other
Drinking	Not at all, rarely, socially, often, very often, desperately
Smoking	No, yes, sometimes, when drinking, trying to quit
Drug use	Never, sometimes, often
Zodiac sign	Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio, Sagittarius, Capricorn, Aquarius, Pisces
Religion	Atheism, agnosticism, Christianity, Catholicism, Judaism, Islam, Hinduism, Buddhism, other
Cats	Has cat(s), likes cats, dislikes cats
Dogs	Has dog(s), likes dogs, dislikes dogs

behaviors probed were: lying to parents, unruly public behavior, damaging property, stealing something worth less than US$50, stealing from a store, and taking part in a fight. Each behavior is recorded simply as present or absent, so there are only two possible responses for each survey item. The data set is included in full in a table within the paper by Li et al.

S4.7 Online dating

This data set contains dating profiles for 59 946 users of the online dating service OKCupid, as posted at `https://www.kaggle.com/datasets/subhamyadav580/dating-site` by Shubham Yadav. The full data set contains both textual descriptions contributed by the users and answers to multiple-choice profile questions, but we focus on the latter only, and on 14 questions in particular, which are listed in Table 3. We include a "missing data" option as an additional possible response for each question, since significant numbers of people did not answer all questions.

References

Bortz AB, Kalos MH, Lebowitz JL (1975) A new algorithm for Monte Carlo simulation of ising spin systems. Journal of Computational Physics 17:10–18

Collins LM, Lanza ST (2010) Latent Class and Latent Transition Analysis. John Wiley and Sons, Hoboken, NJ

Li Y, Lord-Bessen J, Shiyko M, et al (2018) Bayesian latent class analysis tutorial. Multivariate Behavioral Research 53:430–451

Newman MEJ, Barkema GT (1999) Monte Carlo Methods in Statistical Physics. Oxford University Press, Oxford

Walsh CD (2006) Latent class analysis identification of syndromes in Alzheimer's disease: A Bayesian approach. Advances in Methodology and Statistics 3:147–162