CSC 2541: Bayesian Methods for Machine Learning

Radford M. Neal, University of Toronto, 2011

Lecture 4

Problem: Density Estimation

We have observed data, $y_1, \ldots, y_n$, drawn independently from some unknown distribution, whose density we wish to estimate. The observations y_i may be multidimensional.

Possible approaches:

- Use a simple parametric model — eg, multivariate normal.
- Use a non-model-based method — eg, kernel density estimation.
- Use a flexible model for the density — eg, log-spline density model, mixtures.

Problems:

- Densities must be non-negative, and integrate to one.
- For high dimensional data, strong prior assumptions are needed to get good results with a reasonable amount of data.

Problem: Latent Class Analysis

We have multivariate data from a population we think consists of several sub-populations. For example:

- Teachers with different instructional styles.
- Different species of iris.

We don't know which data points came from which sub-populations, or even how many sub-populations there are.

We think that some of the dependencies among the variables are explained by membership in these sub-populations.

We wish to reconstruct the sub-populations ("latent classes") from the dependencies in the observed data.

Mixtures of Simple Distributions

Mixtures of simple distributions are suitable models for both density estimation and latent class analysis. The density of y has the form:

$$\sum_{c=1}^{K} \rho_c \, f(y|\phi_c)$$

The ρ_c are the mixing proportions. The ϕ_c parameterize the simple component densities (in which, for example, the components making up a multidimensional y might be independent).

Some advantages:

- Mixture models produce valid densities.
- With enough components, a mixture can approximate any distribution well.
- Mixtures of simple components are restricted enough to work in many dimensions.
- The mixture components can be interpreted as representing latent classes.

Bayesian Mixture Models

A Bayesian mixture models requires a prior for the mixing proportions, ρ_c, and component parameters, ϕ_c.

We can use a symmetric Dirichlet prior for the ρ_c, with density

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{c=1}^{K} \rho_c^{(\alpha/K)-1} \quad (\rho_c \geq 0,\ \sum_c \rho_c = 1)$$

When α is large, the ρ_c tend to be nearly equal; when α is close to zero, a few of the ρ_c are much bigger than the others.

We will make the ϕ_c be independent under the prior, all with the same distribution, G_0.

There may be higher levels to the model (eg, a prior for α), but let's ignore that possibility for now.

The Model Using Class Indicators

We can express the mixture model using latent variables, c_i, that identify the mixture component (latent class) of each y_i:

$$
\begin{aligned}
y_i \mid c_i,\, \phi &\sim F(\phi_{c_i}) \\
c_i \mid \rho_1, \ldots, \rho_K &\sim \text{Discrete}\,(\rho_1, \ldots, \rho_K) \\
\phi_c &\sim G_0 \\
\rho_1, \ldots, \rho_K &\sim \text{Dirichlet}\,(\alpha/K, \ldots, \alpha/K)
\end{aligned}
$$

The class indicators will have values $1, \ldots, K$.

The model is not "identifiable", since relabelling the classes changes nothing, but this causes no problems — all that really matters is how the class indicators partition the data set (ie, for each i and j, whether $c_i = c_j$ or not).

The Prior After Integrating Out the Mixing Proportions

The mixing proportions (ρ_c) can be eliminated by integrating with respect to their Dirichlet prior. The resulting successive conditional probabilities follow the well-known "law of succession":

$$P(c_i = c \mid c_1, \ldots, c_{i-1}) \;=\; \frac{n_{i,c} + \alpha/K}{i - 1 + \alpha}$$

where $n_{i,c}$ is the number of c_j for $j < i$ that are equal to c.

We could generate from the prior distribution for the c_i and y_i as follows:

- Generate c_1, c_2, …, c_n using the above probabilities (note that $P(c_1 = c) \;=\; 1/K$).
- Generate ϕ_c for $c = 1, \ldots, K$ from G_0.
- Generate each y_i from $F(\phi_{c_i})$, independently.

Exchangeability

Consider a Bayesian model for data items $y_1, \ldots, y_n$ that, given values for the parameters of the model, are independent and identically distributed.

In the *unconditional* distribution of $y_1, \ldots, y_n$, the data items are *not* independent. However, integrating over the parameters of the model shows that the unconditonal distribution of the data items is *exchangeable* — for any permutation π of $1, \ldots, n$,

$$P(Y_1 = y_1, \ldots, Y_n = y_n) \quad = \quad P(Y_1 = y_{\pi(1)}, \ldots, Y_n = y_{\pi(n)})$$

The converse is also true: *de Finetti's representation theorem* says that if the distribution of $y_1, \ldots, y_n$ is exchangeable for all n, it must be expressible as

$$P(y_1, \ldots, y_n) \quad = \quad \int P(\theta) \prod_{i=1}^{n} P(y_i|\theta)\, d\theta$$

For some parameter θ (perhaps infinite-dimensional), some prior $P(\theta)$, and some data distribution $P(y_i|\theta)$, which is the same for all i.

Using Exchangeability for Mixture Models

Due to exchangeability, we can imagine that any particular y_i (along with the corresponding c_i) is the *last* data item.

In particular, from the "law of succession" when probabilities have a Dirichlet prior, we obtain

$$P(c_i = c \mid c_{-i}) \;=\; \frac{n_{-i,c} + \alpha/K}{n - 1 + \alpha}$$

where c_{-i} represents all c_j for $j \neq i$, and $n_{-i,c}$ is the number of c_j for $j \neq i$ that are equal to c.

I will call this the "conditional prior" for c_i.

Gibbs Sampling

We can apply Gibbs sampling to the posterior distribution of this model.

The y_i are known. The state of the Markov chain consists of c_i for $i = 1, \ldots, n$ and ϕ_c for $c = 1, \ldots, K$ (recall we integrated away the ρ_i).

We start from some initial state (eg, with all $c_i = 1$) and then alternately draw each ϕ_c and each c_i from their conditional distributions:

- $\phi_c \mid c_1, \ldots, c_n,\ y_1, \ldots, y_n$

  The conditional distribution for the parameters of one of the component distributions, given the values y_i for which $c_i = c$. (This will be tractable if the prior for ϕ_c is conjugate to the distributional form of the mixture component.)

- $c_i \mid c_{-i},\ \phi_1, \ldots, \phi_K,\ y_i$

  The conditional distribution for one c_i, given the other c_j for $j \neq i$, the parameters of all the mixture components, and the observed value of this data item, y_i.

Gibbs Sampling Updates for the Class Indicators

To pick a new value for c_i during Gibbs sampling, we need to sample from the distribution $c_i \mid c_{-i},\ \phi_1, \ldots, \phi_K,\ y_i$.

This distribution comes from the conditional prior, $c_i \,|\, c_{-i}$, and the likelihood, $f(y_i, \phi_c)$:

$$P(c_i = c \mid c_{-i},\ \phi_1, \ldots, \phi_K,\ y_i) \ = \ \frac{1}{Z}\ \frac{n_{-i,c} + \alpha/K}{n - 1 + \alpha}\ f(y_i, \phi_c)$$

where Z is the required normalizing constant.

It's easy to sample from this distribution, by explicitly computing the probabilities for all K possible value of c_i.

How Many Components?

How many components (K) should we include in our model?

If we set K too small, we won't be able to model the density well. And if we're looking for latent classes, we'll miss some.

If we use a large K, the model will overfit if we set parameters by maximum likelihood. With some priors, a Bayesian model with large K may underfit.

A large amount of research has been done on choosing K, by Bayesian and non-Bayesian methods.

But does choosing K actually make sense?

Is there a better way?

Letting the Number of Components Go to Infinity

For density estimation, there is often reason to think that approximating the real distribution arbitrarily well is only possible as K goes to infinity.

For latent class analysis, there is often reason to think the real number of latent classes is effectively infinite.

If that's what we believe, why not let K be infinite? What happens?

The limiting form of the "law of succession" is

$$P(c_i = c \mid c_1, \ldots, c_{i-1}) \quad = \quad \frac{n_{i,c} + \alpha/K}{i - 1 + \alpha} \quad \rightarrow \quad \frac{n_{i,c}}{i - 1 + \alpha}$$

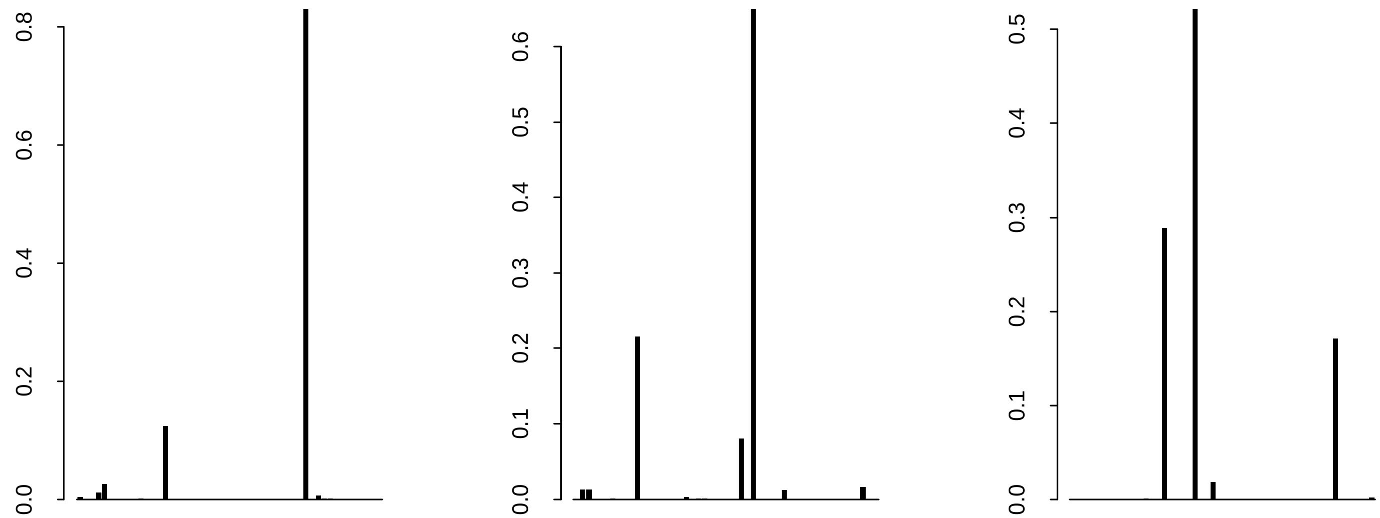
$$P(c_i \neq c_j \text{ for all } j < i \mid c_1, \ldots, c_{i-1}) \quad \rightarrow \quad \frac{\alpha}{i - 1 + \alpha}$$

So even with infinite K, behaviour is reasonable: The probability of the next data item being associated with a new mixture component is neither 0 nor 1.

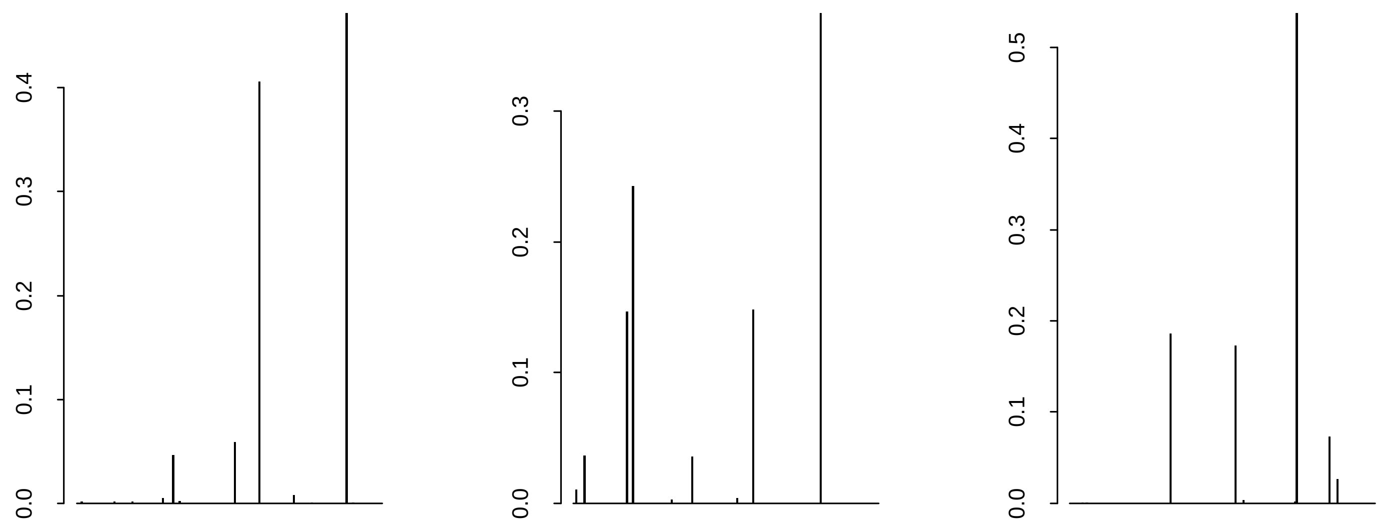
The Prior for Mixing Proportions as K Increases

Three random values from priors for $\rho_1, \ldots, \rho_K$:

$\alpha = 1, \quad K = 50:$



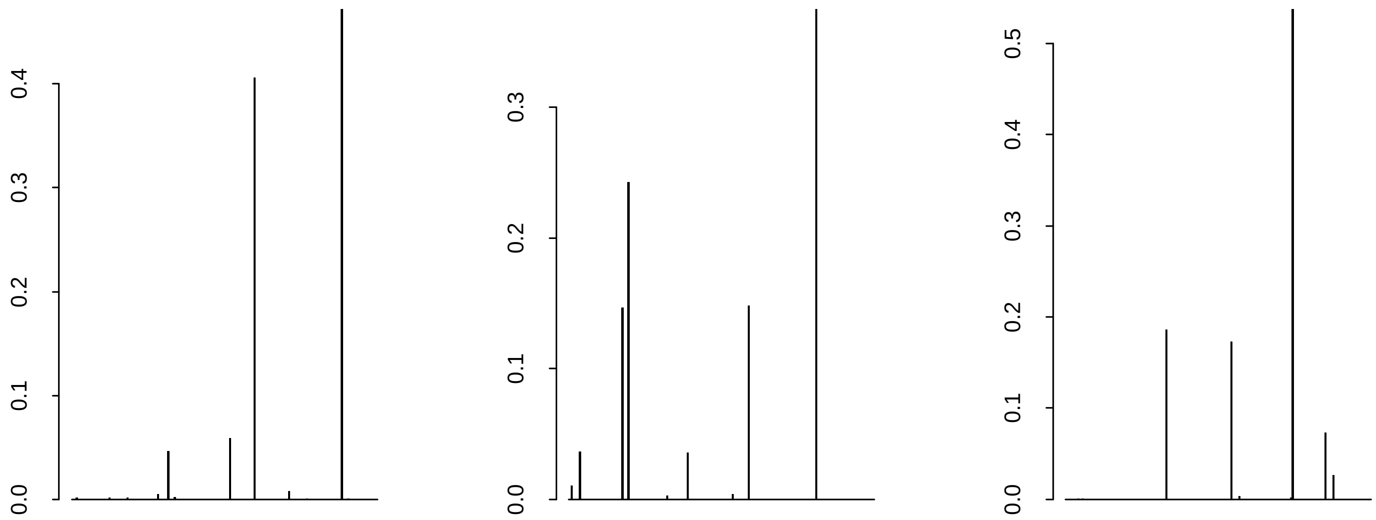
$\alpha = 1, \quad K = 150:$



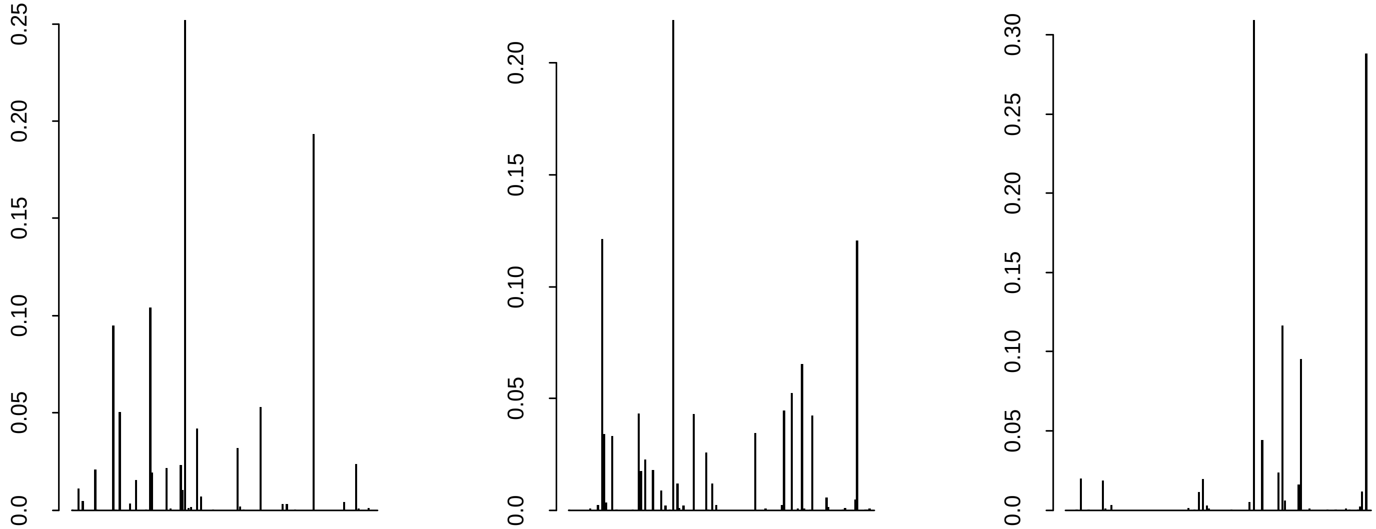
The Prior for Mixing Proportions as α Varies

Three random values from priors for $\rho_1, \ldots, \rho_K$:

$\alpha = 1, \quad K = 150:$



$\alpha = 5, \quad K = 150:$



The Dirichlet Process View

Let $\theta_i = \phi_{c_i}$ be the parameters of the distribution from which y_i was drawn.

When $K \to \infty$, the "law of succession" for the c_i together with the prior (G_0) for the ϕ_c lead to conditional distributions for the θ_i as follows:

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1}$$

$$\sim \quad \frac{1}{i-1+\alpha} \sum_{j<i} \delta(\theta_j) \; + \; \frac{\alpha}{i-1+\alpha} G_0$$

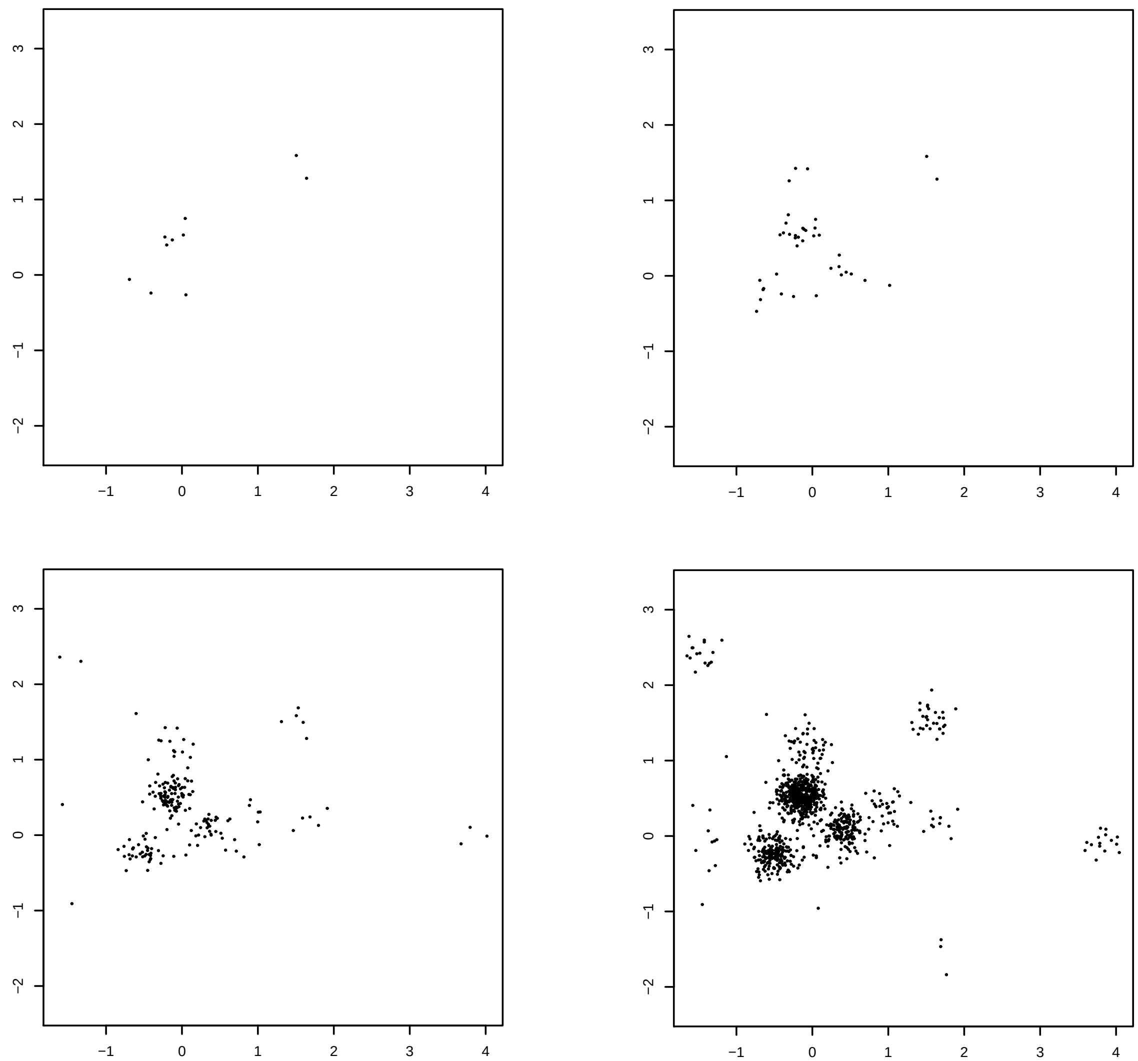
This is the "Polya urn" representation of the *Dirichlet process*, $D(G_0, \alpha)$ — which is a distribution over distributions. We can write the model with infinite K as

$$\begin{aligned} y_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim D(G_0, \alpha) \end{aligned}$$

The name "Dirichlet process mixture model" comes from this view, in which the mixing distribution G has a Dirichlet process as its prior.

Data From a Dirichlet Process Mixture

Data sets of increasing size from a Dirichlet process mixture model with 2D Gaussian distributions for each cluster, with $\alpha = 1$:



$n = 10$ $n = 40$

$n = 200$ $n = 1000$

Can We Do Gibbs Sampling With an Infinite Number of Components?

What becomes of our Gibbs sampling algorithm as $K \to \infty$?

Sampling from $\phi_c \mid c_1, \ldots, c_n,\ y_1, \ldots, y_n$ continues as before, for $c \in \{c_1, \ldots, c_n\}$.

For all other c, the result would be a draw from the prior, G_0. We imagine this having happened, but we don't actually do the infinite amount of work it would require.

To sample from $c_i \mid c_{-i},\ \phi,\ y_i$, we can start by explicitly computing

$$\frac{n_{-i,c}}{n - 1 + \alpha}\ f(y_i, \phi_c)$$

for $c \in c_{-i}$. There are also an infinite number of other possible values for c_i. To do Gibbs sampling, we need to handle them with a finite amount of work.

Gibbs Sampling for Infinite Mixture Models with Conjugate Priors

Consider Gibbs sampling for c_i when K is very large, but finite.

At most n of the classes will be associated with data items (at most $n-1$ with data items other than the ith). The probability of setting c_i to such a c is proportional to

$$\frac{n_{-i,c} + \alpha/K}{n-1+\alpha}\ f(y_i, \phi_c)$$

For any value of c not in c_{-i}, the product of the conditional prior and the likelihood will be

$$\frac{\alpha/K}{n-1+\alpha}\ f(y_i, \phi_c)$$

where the ϕ_c are drawn from the prior, G_0. As $K \to \infty$, the *total* probability of setting c_i to *any* c not in c_{-i} is proportional to

$$\frac{\alpha}{n-1+\alpha}\ \int f(y_i, \phi)\, dG_0(\phi)$$

More on Gibbs Sampling for Infinite Mixture Models with Conjugate Priors

If G_0 is a conjugate prior for F, we can evaluate $\int f(y_i, \phi)\, dG_0(\phi)$ analytically.

We can then figure out the correct probability for setting c_i to be any of the other c_j, or any of the infinity number of c that are not currently in use. Specifically:

$$
\begin{aligned}
P(c_i = c_j \mid c_{-i},\ \phi,\ y_i) \quad &= \quad \frac{1}{Z}\, \frac{n_{-i,c_j}}{n-1+\alpha}\, f(y_i, \phi_{c_j}), \quad \text{for } j \in c_{-i} \\
P(c_i \neq c_j \text{ for all } j \in c_{-i} \mid c_{-i},\ \phi,\ y_i) \quad &= \quad \frac{1}{Z}\, \frac{\alpha}{n-1+\alpha} \int f(y_i, \phi)\, dG_0(\phi)
\end{aligned}
$$

where Z is the appropriate normalizing constant, the same for both equations.

If we choose a previously unused c for c_i, we also explicitly choose a value for ϕ_c, from the posterior for ϕ given the prior G_0 and the single data point y_i.

When a previously used c is no longer used, we stop representing it explicitly, forgetting the corresponding ϕ_c. (If we kept it around, it would never be used again, since for such a c, $n_{-i,c}$ will always be zero.)

The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm generalizes the Metropolis algorithm to allow for non-symmetric proposal distributions.

When sampling from $\pi(x)$, a transition from state x to state x' goes as follows:

1) A "candidate", x^*, is proposed according to some probabilities $S(x, x^*)$, not necessarily symmetric.

2) This candidate, x^*, is accepted as the next state with probability

$$\min\left[1,\ \frac{\pi(x^*)S(x^*, x)}{\pi(x)S(x, x^*)}\right]$$

   If x^* is accepted, then $x' = x^*$. If x^* is instead rejected, then $x' = x$.

One can easily show that transitions defined in this way satisfy detailed balance, and hence leave π invariant.

A Metropolis-Hastings Algorithm for Infinite Mixture Models with Non-Conjugate Priors

We can update c_i using a M-H proposal from the conditional prior, which for finite K is

$$P(c_i = c^* \mid c_{-i}) \;=\; \frac{n_{-i,c^*} + \alpha/K}{n-1+\alpha} \;=\; S(c, c^*)$$

We need to accept or reject so as to leave invariant the conditional distribution for c_i:

$$\begin{aligned} \pi(c) \;&=\; P(c_i = c \mid c_{-i},\ \phi_1, \ldots, \phi_K,\ y_i) \\ &=\; \frac{1}{Z}\,\frac{n_{-i,c} + \alpha/K}{n-1+\alpha}\; f(y_i, \phi_c) \end{aligned}$$

The required acceptance probability for changing c_i from c to c^* is

$$\min\left[1,\ \frac{\pi(c^*)S(c^*, c)}{\pi(c)S(c, c^*)}\right] \;=\; \min\left[1,\ \frac{f(y_i, \phi_{c^*})}{f(y_i, \phi_c)}\right]$$

When $K \to \infty$, we can still sample from the conditional prior. If we pick a c^* that is a currently unused, we pick a value of ϕ_{c^*} to go with it from G_0. The acceptance probability is then easily computed.