Mixture modelling notes

Paul D. W. Kirk[1,*]

[1]MRC Biostatistics Unit, Cambridge, UK
[*]paul.kirk@mrc-bsu.cam.ac.uk

1 Mixture modelling

Suppose we have data consisting of n observations, $D = \{\mathbf{v}_i\}_{i=1}^n$.

We model the data using a mixture model with K components (where K could be finite or infinite), as follows:

$$p(\mathbf{v}|\rho,\pi) = \sum_{k=1}^{K} \pi_k f_{\mathbf{v}}(\mathbf{v}|\rho_k), \tag{1}$$

where π_k is the mixture weight associated with the k-th component, and ρ_k denotes the parameters associated with the k-th component.

As is common for mixture models, we introduce latent component allocation variables, c_i, where $c_i = k$ if the i-th observation $\mathbf{v}_i$ is associated with the k-th component, and $p(c_i = k|\pi) = \pi_k$. Then,

$$p(\mathbf{v}_i|c_i,\rho) = f_{\mathbf{v}}(\mathbf{v}_i|\rho_{c_i}), \tag{2}$$

and hence

$$\begin{aligned} p(\mathbf{v}_i, c_i = k|\rho,\pi) &= f_{\mathbf{v}}(\mathbf{v}_i|\rho_k)p(c_i = k|\pi) && (3)\\ &= f_{\mathbf{v}}(\mathbf{v}_i|\rho_k)\pi_k. && (4) \end{aligned}$$

Integrating out c_i by summing over all K possible values, we obtain (as we would hope):

$$p(\mathbf{v}_i|\rho,\pi) = \sum_{k=1}^{K} \pi_k f_{\mathbf{v}}(\mathbf{v}_i|\rho_k). \tag{5}$$

Making the usual conditional independence assumptions, the full joint model for $\mathbf{v}_i, c_i, \rho, \pi$ is:

$$\begin{aligned} p(\mathbf{v}_i, c_i, \rho, \pi) &= f_{\mathbf{v}}(\mathbf{v}_i|\rho_{c_i})p(c_i|\pi)p(\pi)p(\rho) && (6)\\ &= f_{\mathbf{v}}(\mathbf{v}_i|\rho_{c_i})p(c_i|\pi)p(\pi)\prod_{k=1}^{K} p(\rho_k), && (7) \end{aligned}$$

where we assume independent priors for the component-specific parameters, ρ_k.

For the full dataset, we have:

$$\begin{aligned} p(\mathbf{v}_1,\ldots,\mathbf{v}_n,c_1,\ldots,c_n,\rho,\pi) &= \left(\prod_{i=1}^{n} f_{\mathbf{v}}(\mathbf{v}_i|\rho_{c_i})p(c_i|\pi)\right) p(\pi)p(\rho) && (8)\\ &= \left(\prod_{i=1}^{n} f_{\mathbf{v}}(\mathbf{v}_i|\rho_{c_i})p(c_i|\pi)\right) p(\pi)\prod_{k=1}^{K} p(\rho_k). && (9) \end{aligned}$$

1.1 Conditionals for Gibbs sampling (finite K case)

Given Equation (9), it is straightforward to write down the conditionals for Gibbs sampling. For the time being, we assume finite K (from which we will later derive the infinite limit).

1.1.1 Conditional for ρ_k

By examination of the RHS of Equation 9, we have:

$$p(\rho_k|\mathbf{v}_1,\dots,\mathbf{v}_n,c_1,\dots,c_n,\rho_{-k},\pi) \propto p(\rho_k)\prod_{i:c_i=k} f_{\mathbf{v}}(\mathbf{v}_i|\rho_{c_i}), \tag{10}$$

where ρ_{-k} denotes the set comprising all ρ_j for which $j \neq k$. Thus the conditional for ρ_k is the posterior density for ρ_k given all $\mathbf{v}_i$ for which $c_i = k$. Note that if there are no $\mathbf{v}_i$ for which $c_i = k$ (i.e. if the k-th component has no observations associated with it), then ρ_k is simply sampled from the prior, $p(\rho_k)$.

1.1.2 Conditional for π

By examination of the RHS of Equation 9, we have:

$$p(\pi|\mathbf{v}_1,\dots,\mathbf{v}_n,c_1,\dots,c_n,\rho,\pi) \propto \left(\prod_{i=1}^{n} p(c_i|\pi)\right) p(\pi). \tag{11}$$

$$\tag{12}$$

Hence, the conditional for π is the posterior for π given the values taken by the categorical latent allocation variables, c_i, $i = 1,\dots,n$. If we take a conjugate Dirichlet prior, this posterior is available in closed form.

1.1.3 Conditional for c_i

By examination of the RHS of Equation 9, we have:

$$p(c_i = k|\mathbf{v}_1,\dots,\mathbf{v}_n,\rho,\pi,c_{-i}) \propto p(c_i = k|\pi) f_{\mathbf{v}}(\mathbf{v}_i|\rho_k), \tag{13}$$

$$= \pi_k f_{\mathbf{v}}(\mathbf{v}_i|\rho_k), \tag{14}$$

where c_{-i} denotes the set comprising all c_j for which $j \neq i$. Since $\sum_{k=1}^{K} p(c_i = k|\mathbf{v}_1,\dots,\mathbf{v}_n,\rho,\pi,c_{-i}) = 1$, it follows that the conditional is:

$$p(c_i = k|\mathbf{v}_1,\dots,\mathbf{v}_n,\rho,\pi,c_{-i}) = \frac{\pi_k f_{\mathbf{v}}(\mathbf{v}_i|\rho_k)}{\sum_{k=1}^{K} \pi_k f_{\mathbf{v}}(\mathbf{v}_i|\rho_k)}, \tag{15}$$

which may be straightforwardly evaluated for finite K.

1.1.4 Marginalising π

Taking a conjugate Dirichlet prior for π, an alternative strategy is to marginalise π rather than to sample it. We assume a symmetirc Dirichlet prior with concentration parameter α/K.

Note that the c_i's are only conditionally independent of one another given π, so if we marginalise π then we must be careful to model the dependence of c_i on c_{-i} in our conditional for c_i.

We have

$$p(c_i = k|\mathbf{v}_1,\dots,\mathbf{v}_n,\rho,\pi,c_{-i},\alpha) \propto p(c_i = k|c_{-i},\pi,\alpha) f_{\mathbf{v}}(\mathbf{v}_i|\rho_k) \qquad \text{[cf. Equation (13)]}. \tag{16}$$

To marginalise π, we must therefore evaluate $\int_\pi p(c_i = k|c_{-i},\pi,\alpha)p(\pi|\alpha)d\pi = p(c_i = k|c_{-i},\alpha)$, which is the conditional prior for c_i given the values for the other latent allocation variables, c_{-i}.

We have,

$$p(c_i = k|c_{-i},\alpha) = \int_\pi p(c_i = k|c_{-i},\pi,\alpha)p(\pi|\alpha)d\pi \tag{17}$$

$$= \int_\pi \frac{p(c_i = k,c_{-i}|\pi,\alpha)}{p(c_{-i}|\pi,\alpha)} p(\pi|\alpha)d\pi \tag{18}$$

$$= \frac{\int_\pi p(c_i = k,c_{-i}|\pi)p(\pi|\alpha)d\pi}{\int_\pi p(c_{-i}|\pi)p(\pi|\alpha)d\pi}, \tag{19}$$

where in the final line we exploit the fact that the c_i's are conditionally independent of α, given π.

In order to proceed, we must evaluate this fraction. To do this we require a standard result about Dirichlet distributions, which says that moments of random variables distributed according to a symmetric Dirichlet distribution with concentration parameter α/K can be expressed as follows:

$$E\left[\prod_{k=1}^{K}\pi_k^{m_k}\right] = \frac{\Gamma(\sum_{k=1}^{K}(\alpha/K))}{\Gamma(\sum_{k=1}^{K}((\alpha/K)+m_k))} \times \prod_{k=1}^{K}\frac{\Gamma((\alpha/K)+m_k)}{\Gamma(\alpha/K)}, \tag{20}$$

where the m_k's are any natural numbers.

Moreover, we note the following two equalities:

$$p(c_i = k, c_{-i}|\pi) = \pi_k^{n_{-i,k}+1}\prod_{\substack{c=1,\dots,K\\c\neq k}}\pi_c^{n_{-i,c}},$$

and

$$p(c_{-i}|\pi) = \pi_k^{n_{-i,k}}\prod_{\substack{c=1,\dots,K\\c\neq k}}\pi_c^{n_{-i,c}},$$

where $n_{-i,c}$ is the number of c_j's with $j \neq i$ for which $c_j = c$. It then follows that we may use the result given in Equation (20) in order to evaluate the numerator and denominator in the RHS of Equation (19). After some algebra, and exploiting the property of Gamma functions that $\Gamma(t+1) = t\Gamma(t)$, we obtain:

$$p(c_i = k|c_{-i}, \alpha) = \frac{n_{-i,k}+\alpha/K}{n-1+\alpha}. \tag{21}$$

Hence,

$$p(c_i = k|\mathbf{v}_1, \dots, \mathbf{v}_n, \rho, c_{-i}, \alpha) \propto \frac{n_{-i,k}+\alpha/K}{n-1+\alpha} \times f_{\mathbf{v}}(\mathbf{v}_i|\rho_k). \tag{22}$$

Moreover, since K is finite, we may straightforwardly evaluate the equality:

$$p(c_i = k|\mathbf{v}_1, \dots, \mathbf{v}_n, \rho, c_{-i}, \alpha) = \frac{1}{Z}\frac{n_{-i,k}+\alpha/K}{n-1+\alpha} \times f_{\mathbf{v}}(\mathbf{v}_i|\rho_k), \tag{23}$$

where

$$Z = \sum_{c=1}^{K}\left(\frac{n_{-i,c}+\alpha/K}{n-1+\alpha} \times f_{\mathbf{v}}(\mathbf{v}_i|\rho_c)\right). \tag{24}$$

1.1.5 Marginalising ρ

Similarly, if a conjugate prior is available for the ρ_k's, then these may be marginalised too. **Note** that (similar to the case with the c_i's when we marginalised π) the $\mathbf{v}_i$'s are only conditionally independent of one another given the ρ_k's and the c_i's, so if we marginalise ρ then we must be careful to model the dependence of $\mathbf{v}_i$ on $\mathbf{v}_{-i}$ in our conditional for c_i.

After some algebra, it is straightforward to show that marginalising ρ gives the following for the conditional for c_i:

$$p(c_i = k|\mathbf{v}_1, \dots, \mathbf{v}_n, c_{-i}, \alpha) = \frac{1}{Z}\frac{n_{-i,k}+\alpha/K}{n-1+\alpha} \times \int_{\rho_k} f_{\mathbf{v}}(\mathbf{v}_i|\rho_k)p(\rho_k|\mathbf{v}_{-i,k})d\rho_k, \tag{25}$$

where $\mathbf{v}_{-i,k}$ denotes all observations $\mathbf{v}_j$ for which $j \neq i$ and $c_j = k$, and hence $p(\rho_k|\mathbf{v}_{-i,k})$ is the posterior for ρ_k given all of the observations currently associated with component k (excluding $\mathbf{v}_i$). If there are no $\mathbf{v}_j$ for which $j \neq i$

and $c_j = k$ (i.e. if k is a component to which no other observations have been allocated), then we say that the k-th component is *empty* and define $p(\rho_k|\mathbf{v}_{-i,k}) := p(\rho_k)$ to be the prior for ρ_k.

When implementing the sampler, it is useful to observe that

$$p(\rho_k|\mathbf{v}_{-i,k}) = \frac{f_{\mathbf{v}}(\mathbf{v}_{-i,k}|\rho_k)p(\rho_k)}{\int_{\rho_k} f_{\mathbf{v}}(\mathbf{v}_{-i,k}|\rho_k)p(\rho_k)d\rho_k}, \text{ if the } k\text{-th component is not empty.}$$

Hence, still assuming that the k-th component is not empty, the integral in Equation (25) is

$$\int_{\rho_k} f_{\mathbf{v}}(\mathbf{v}_i|\rho_c)p(\rho_k|\mathbf{v}_{-i,k})d\rho_k = \frac{\int_{\rho_k} f_{\mathbf{v}}(\mathbf{v}_i, \mathbf{v}_{-i,k}|\rho_k)p(\rho_k)d\rho_k}{\int_{\rho_k} f_{\mathbf{v}}(\mathbf{v}_{-i,k}|\rho_k)p(\rho_k)d\rho_k}, \tag{26}$$

which is a ratio of marginal likelihoods: one in which we include $\mathbf{v}_i$ amongst the observations associated with component k, and one in which we exclude $\mathbf{v}_i$ from the observations associated with component k.

This expression aids the interpretation of the sampler: at each iteration, and for each component, we weigh the evidence that $\mathbf{v}_i$ is associated with component k against the evidence that $\mathbf{v}_i$ is *not* associated with component k (given the other observations currently associated with that component, $\mathbf{v}_{-i,k}$). Intuitively, this expression ensures that we are more likely to allocate $\mathbf{v}_i$ to a component to which similar observations have previously been allocated.

Note also that the term $\frac{n_{-i,k}+\alpha/K}{n-1+\alpha}$ in Equation (25) represents the conditional prior probability that $\mathbf{v}_i$ should be allocated to component k represents our prior belief that we should allocate $\mathbf{v}_i$ to component k, given the allocation of all of the other observations. Since $n_{-i,k}$ is in the numerator, this expresses a "rich-get-richer" prior belief; i.e. that, *a priori*, we are more likely to assign $\mathbf{v}_i$ to a component that already has many observations assigned to it, rather than to one with fewer.

1.1.6 Final note

Note that, having marginalised the π_k's and ρ_k's, we may use Equation (25) to sample just the c_i's, without having to sample any other parameters. The one exception is the α hyperparameter, which we may either fix or sample (using, for example, the approach described in Escobar and West, 1995).

2 Modelling categorical data

We now consider the specific case in which we have categorical covariates.

2.1 Modelling the covariates

We assume that each covariate (i.e. each element of the vector $\mathbf{v}_i$) is categorical, with the j-th covariate having R_j categories, which we label as $1, 2, \ldots, R_j$. We model the data using categorical distributions. We define $\phi_{k,j,r}$ to be the probability that the j-th covariate takes value r in the k-th component, and write $\Phi_{k,j} = [\phi_{k,j,1}, \phi_{k,j,2}, \ldots, \phi_{k,j,R_j}]$ for the collection of probabilities associated with the j-th covariate in the k-th component. We further define $\Phi_k = \{\Phi_{k,1}, \Phi_{k,2}, \ldots, \Phi_{k,J}\}$ to be the collection of all probabilities (over all J covariates) associated with the k-th component, and $\Phi = \{\Phi_k\}_{k \in \mathscr{C}}$ to be the set of all Φ_k's that are associated with non-empty components (here, $\mathscr{C} = \{k : c_i = k \text{ for some } i \in \{1, \ldots, n\}\}$).

We assume that the covariates are conditionally independent, given their component allocation, so that

$$\begin{aligned} f_{\mathbf{v}}(\mathbf{v}_i = [v_{i1}, v_{i2}, \ldots, v_{iJ}]|\Phi, c_i = k) &= \phi_{k,j,v_{i1}}\phi_{k,j,v_{i2}}\ldots\phi_{k,j,v_{iJ}} & (27) \\ &= \prod_{j=1}^{J} \phi_{k,j,v_j} & (28) \end{aligned}$$

2.1.1 Conditional for $\Phi_{k,j}$

From Equation (10), the conditional that we require for Gibbs sampling is the posterior for $\Phi_{k,j}$, given the observations associated with the k-th component. For each j, we adopt a conjugate Dirichlet prior for $\Phi_{k,j}$,

$$\Phi_{k,j} \sim \text{Dirichlet}(\mathbf{a}_j),$$

where $\mathbf{a}_j = [a_{j,1},\ldots,a_{j,R_j}]$ is the vector of concentration parameters. The posterior is then:

$$\Phi_{k,j}|v_{i_1,j},v_{i_2,j},\ldots,v_{i_{n_k},j},\mathbf{a}_j \sim \text{Dirichlet}(\mathbf{a}_j + [s_{k,j,1},s_{k,j,2},\ldots,s_{k,j,R_j}]), \tag{29}$$

where $\mathbf{v}_{i_1},\mathbf{v}_{i_2},\ldots,\mathbf{v}_{i_{n_k}}$ are the observations associated with component k, and $s_{k,j,r}$ is defined to be the number of observations associated with component k for which the j-th covariate is in category r.

2.1.2 Marginalising $\Phi_{k,j}$

We may also integrate out $\Phi_{k,j}$ in order to write down the marginal likelihood associated with $v_{i_1,j},v_{i_2,j},\ldots,v_{i_{n_k},j}$. Note that the marginal likelihood is (by definition) the prior expectation of the product $\phi_{k,j,1}^{s_{k,j,1}}\ldots\phi_{k,j,R_j}^{s_{k,j,R_j}}$. We may therefore use the same standard result that was used to derive Equation (20) in order to immediately write down the marginal likelihood. Still assuming that $\mathbf{v}_{i_1},\mathbf{v}_{i_2},\ldots,\mathbf{v}_{i_{n_k}}$ are the observations associated with component k, we have:

$$p(v_{i_1,j},v_{i_2,j},\ldots,v_{i_{n_k},j}|\mathbf{a}_j) = \frac{\Gamma(\sum_{r=1}^{R_j} a_{j,r})}{\Gamma(\sum_{r=1}^{R_j}(a_{j,r}+s_{k,j,r}))} \times \prod_{r=1}^{R_j} \frac{\Gamma(a_{j,r}+s_{k,j,r})}{\Gamma(a_{j,r})}. \tag{30}$$

To shorten notation, define $V_k = \{\mathbf{v}_{i_1},\mathbf{v}_{i_2},\ldots,\mathbf{v}_{i_{n_k}}\}$ to be the set of observations associated with component k, and $V_{k,j} = \{v_{i_1,j},v_{i_2,j},\ldots,v_{i_{n_k},j}\}$ to be the set containing the j-th elements of the vectors in V_k. Since we assume that the covariates are conditionally independent, given their component allocation, it follows that:

$$p(V_k|\mathbf{a}_1,\ldots,\mathbf{a}_J) = \prod_{j=1}^{J} p(V_{k,j}|\mathbf{a}_j), \tag{31}$$

where $p(V_{k,j}|\mathbf{a}_j) = p(v_{i_1,j},v_{i_2,j},\ldots,v_{i_{n_k},j}|\mathbf{a}_j)$ is as given in Equation (30).

2.2 Joint marginal likelihood

In order to proceed, we need an expression for the marginal likelihood associated with V_k.

$$p(V_k|\mathbf{a}_1,\ldots,\mathbf{a}_J,\mathbf{a}_y) = \prod_{j=1}^{J} p(V_{k,j}|\mathbf{a}_j), \tag{32}$$

where the expression for $p(V_{k,j}|\mathbf{a}_j) = p(v_{i_1,j},v_{i_2,j},\ldots,v_{i_{n_k},j}|\mathbf{a}_j)$ is as given in Equation (30). Note that:

1. Setting $V_k = \{\mathbf{v}_i,\mathbf{v}_{-i,k}\}$, we can evaluate the marginal likelihood in the numerator in Equation (**??**).

2. Setting $V_k = \{\mathbf{v}_{-i,k}\}$, we can evaluate the marginal likelihood in the denominator in Equation (**??**).

3. Setting $V_k = \{\mathbf{v}_i\}$, we can evaluate the marginal likelihood in Equation (**??**).

Thus, we may evaluate all of the terms required for the conditionals for c_i, and hence (leaving aside, for the time being, the problem of sampling α) we have everything we need in order to perform inference for our model.