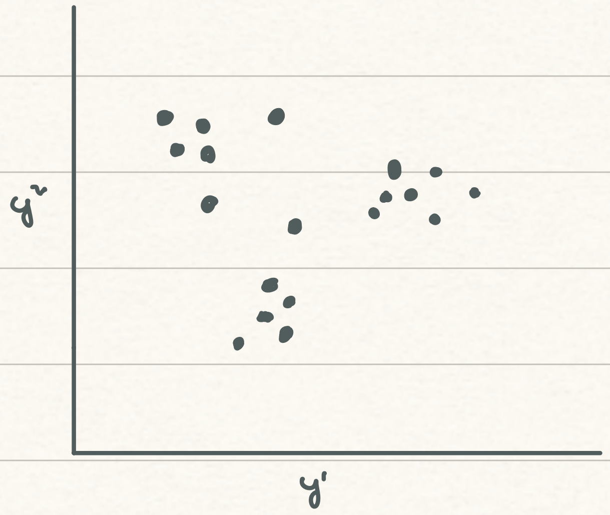
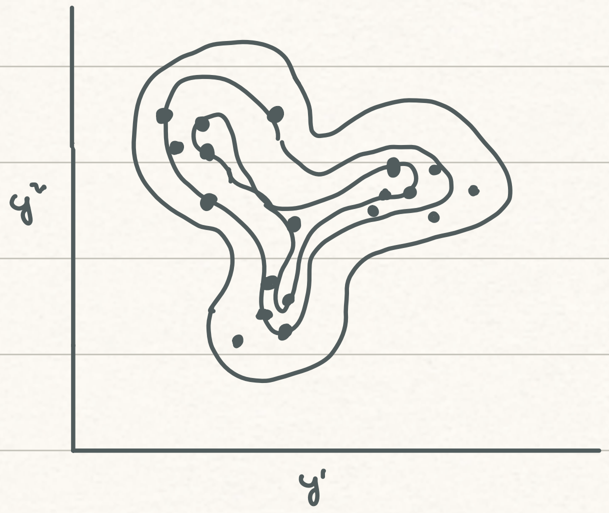
Mixture models

Consider the 2d data below, which shows signs of **clustering**.



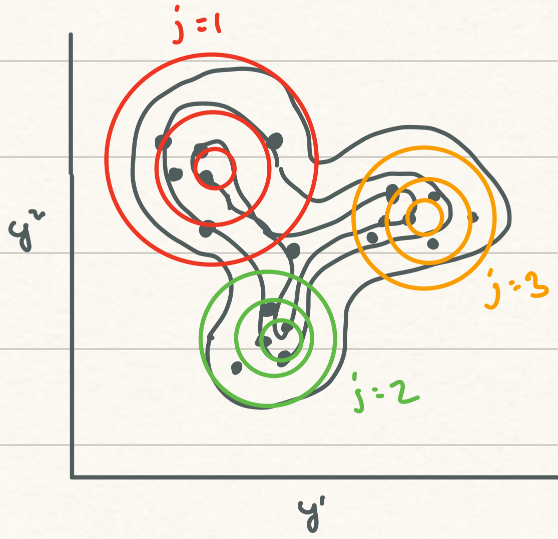
$$y_i = \begin{bmatrix} y_i^1 \\ y_i^2 \end{bmatrix}, \quad i = 1 \ldots n$$

Where might a new data point arrive? The **marginal likelihood** might be complicated:



$$p(y_i)$$

We can instead represent this as a **mixture distribution** of simpler components:



$$p(y_i) = \sum_{j=1}^{K} \pi_j \, p(y_i \mid z_i = j)$$

mixture weights

class-conditional likelihood

Simple Gaussian mixture model

Let's build a simple model for the data above. Assume that each class-conditional likelihood is an isotropic Gaussian:

$$p(y_i \mid z_i = j) = \mathcal{N}(y_i \mid \mu_j, I\sigma_0^2)$$

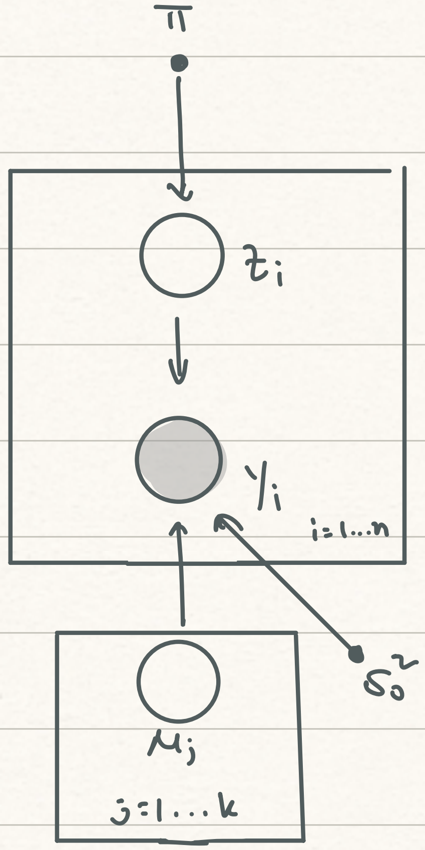
$$p(z_i = j) = \pi_j$$

Each class (or component) is associated with a mean parameter, in this case. (We'll assume that the variance is a fixed hyperparameter shared across all classes).

We can place a (conditionally conjugate) Gaussian prior over the unknown means:

$$p(\mu_j) = \mathcal{N}(\mu_j \mid m_0, I\lambda_0^{-1})$$

Here's the graphical model:



Assume for now that the mixture weights are known, and the only unknown are the class means. (We will relax this assumption later). What is the posterior?

$$p(\mu_{1:k} \mid y_{1:n}) \propto p(\mu_{1:k}, y_{1:n})$$

$$= \underbrace{\sum_{z_1=1}^{k} \cdots \sum_{z_n=1}^{k}}_{} p(\mu_{1:k}, y_{1:n}, z_{1:n})$$

Marginalizing over the **cluster assignments** involves summing over K^n terms

$$\text{e.g. } K=3,\ n=200 \Rightarrow 3^{200}$$

this is much more than all atoms in the universe

In general, the **model evidence** under even a simple mixture model with K=3 classes is **intractable**, and therefore the posterior is too. Nevertheless, we will be able to exploit the modular structure of the model to fit it tractably.

Complete data likelihood and conditional posterior

We can think about the cluster assignments as **missing data**, since there is one of these variables per data point. The "complete data likelihood" is then:

$$p(y_{1:n}, z_{1:n} \mid \mu_{1:k}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \mathcal{N}(y_i \mid \mu_k, \sigma_0^2 I)^{1(z_i = j)}$$

It will be convenient to alias the indicator functions as:

$$f_{ij} \equiv 1(z_j = j)$$

Now what its the posterior of the means given the complete data (including assignments)?

$$p(\mu \mid y, z) = \frac{p(y, z \mid \mu)\, p(\mu)}{p(y, z)}$$

$$\propto_\mu \prod_j \mathcal{N}(\mu_j \mid m_0, I\lambda_0^{-1}) \prod_i \mathcal{N}(y_i \mid \mu_j, I\sigma_0^2)^{f_{ij}}$$

$$\propto_\mu \prod_j \exp\left(-\frac{\lambda_0}{2}(m_0 - \mu_j)^2\right) \prod_i \exp\left(-\frac{1}{2\sigma_0^2}(\mu_j - y_i)^2\right)^{f_{ij}}$$

$$\propto_\mu \prod_j \exp\left(-\frac{\lambda_0}{2}\left[\mu_j^2 - 2\mu_j m_0\right]\right) \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n} f_{ij}\left[\mu_j^2 - 2\mu_j y_i\right]\right)$$

$$\propto_\mu \prod_{j=1}^{k} \exp\left(-\frac{1}{2}\mu_j^2 \underbrace{\left[\lambda_0 + \frac{1}{\sigma_0^2}\sum_{i=1}^{n} f_{ij}\right]}_{\equiv \lambda_j} + \mu_j \underbrace{\left[m_0 + \frac{1}{\sigma_0^2}\sum_{i=1}^{n} f_{ij} y_i\right]}_{= \lambda_j m_j}\right)$$

We recognize this as a product of Gaussian kernels, where the posterior parameters involve sums over the indicators.

$$\propto \prod_j \mathcal{N}(\mu_j \mid m_j, \lambda_j^{-1})$$

Cluster "responsibilities"

So if we know the cluster assignments, things become easy / tractable. If the complete conditional of the cluster assignments is tractable, then we could flip back-and-forth between updating both (e.g., Gibbs sampling). The complete conditional of Zi is:

$$P(z_i = j \mid y, z_{\setminus i}, \mu) \underset{z_i}{\propto} P(y, z_{\setminus i}, \mu, z_i = j)$$

$$\propto P(y_i \mid z_i = j, \mu)\, P(z_i = j)$$

Notice that it does not depend on the other Zi's.

$$\propto \mathcal{N}(y_i \mid \mu_j, \sigma_0^2 I)\, \pi_j$$

The support of Zi is finite/discrete {1…K}, so the posterior obtains with a simple sum:

$$P(z_i = j \mid y_i, \mu) = \frac{\mathcal{N}(y_i \mid \mu_j, \sigma_0^2 I)\, \pi_j}{\sum_{j'=1}^{K} \mathcal{N}(y_i \mid \mu_{j'}, \sigma_0^2 I)\, \pi_{j'}}$$

This conditional plays an important role in mixture-modeling and is often called the "responsibility of cluster j to data point i":

$$= r_{ij} = \mathbb{E}\left[\delta_{ij} \mid y_i, \mu_{1:K}\right]$$

It will also be convenient at times to think of the vector of responsibilities for i:

$$\vec{r}_i = \begin{bmatrix} r_{i1} \\ \vdots \\ r_{iK} \end{bmatrix} \in \Delta_K$$

K-dimensional simplex

$$= \mathbb{E}\left[\vec{\delta}_i \mid y_i, \mu_{1:K}\right]$$

$$\text{where } \vec{\delta}_i = \begin{bmatrix} \delta_{i1} \\ \vdots \\ \delta_{iK} \end{bmatrix}$$

Possible algorithms: interated conditional modes (ICM) and Gibbs:

With the two complete conditionals above, we could run one of the following:

Iterated conditional models:

```
until convergence:
    for i : 1 ... n:
        $z_i \leftarrow \arg\max_j r_{ij}$
    for j : 1 ... K:
        $\mu_j \leftarrow m_j$
```

This is the **K-means algorithm.**

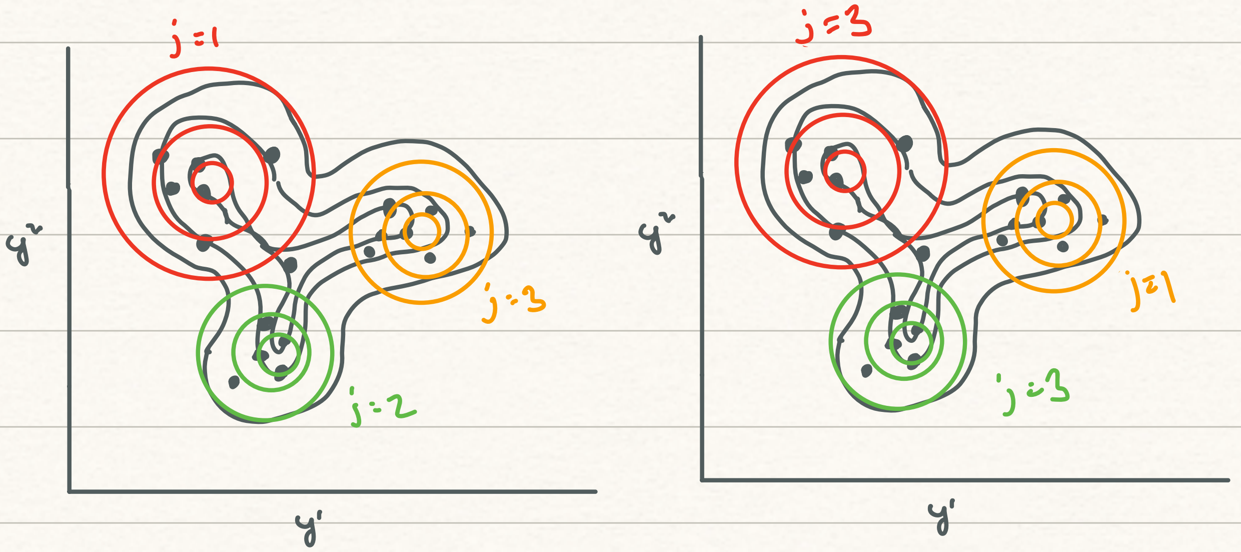
Gibbs sampling

```
for m : 1 ... M:
    for i : 1 ... n:
        $z_i \sim \text{Categorical}(\vec{r}_i)$
    for j : 1 ... K:
        $\mu_j \sim \mathcal{N}(m_j, \lambda_j^{-1} I)$
```

Both of these can work. ICM can be sensitive to initialization, and can lack theoretical guarantees (depending on the model). Gibbs sampling on the other hand is not (nearly as) sensitive to initialization and is theoretically guaranteed to target the full posterior over cluster assignments and parameters. However, the full posterior may be more than we want. Moreover, we will have to contend with the **problem of "label-switching"** (see below). Next class we will explore a third option to fit mixture models: the **EM algorithm.**

Label-switching

Consider the two pictures below, where the K=3 Gaussian components are identical, but their indices have been permitted:



Both of these configurations have non-zero probability under the posterior, and an MCMC chain may thus sample both. This means that **posterior expectations which rely on the cluster indices having a fixed meaning** are not a well-defined.

e.g. $\frac{1}{M}\sum_{m=1}^{M} \mathbb{1}(z_i^m = 3)$ ill-defined ✗ $\frac{1}{M}\sum_{m=1}^{M} \mu^m_{z_i^m}$ well-defined ✓

$\frac{1}{M}\sum_{m=1}^{M} \mu_j^m$ ill-defined ✗ $\frac{1}{M}\sum_{m=1}^{M} \mathbb{1}(z_i = z_{i'})$ well-defined ✓

Categorical vs Multinoulli vs Multinomial

$$z_i \sim \text{Categorical}(\pi), \qquad z_i \in \{1 \dots k\}$$

$$\vec{f}_i \sim \text{Multinoulli}(\pi), \qquad \vec{f}_i \in \left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \right\}$$

$$\equiv \text{Multinomial}(1, \pi)$$

Adding a prior over mixture weights

$$\pi \in \Delta_K$$

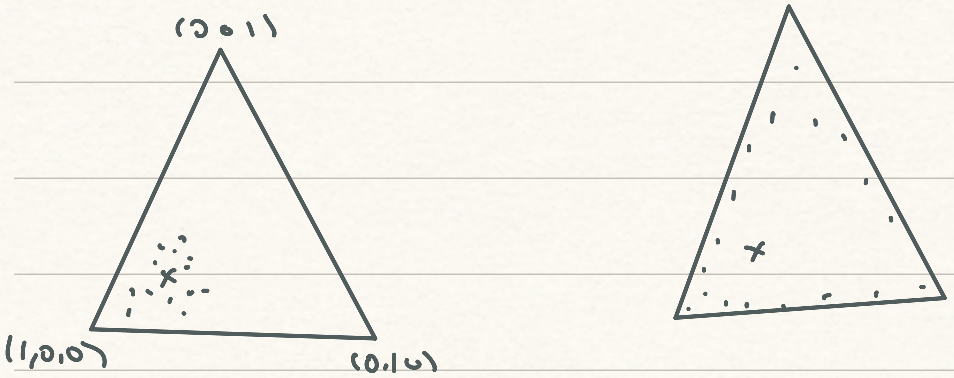
The vector of mixture weights lives on the K-dimensional simplex. What is an appropriate prior? The Dirichlet distribution:

$$\pi \sim \text{Dirichlet}(\alpha), \qquad \alpha = [\alpha_1 \cdots \alpha_K]$$

$$p(\pi \mid \alpha) = \frac{\Gamma(\sum \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_j \pi_j^{\alpha_j - 1} \; \mathbb{1}(\pi \in \Delta_K)$$

$$\mathbb{E}[\pi] = \left[\frac{\alpha_1}{\alpha_0} \cdots \frac{\alpha_K}{\alpha_0} \right], \qquad \alpha_0 = \sum_j \alpha_j$$

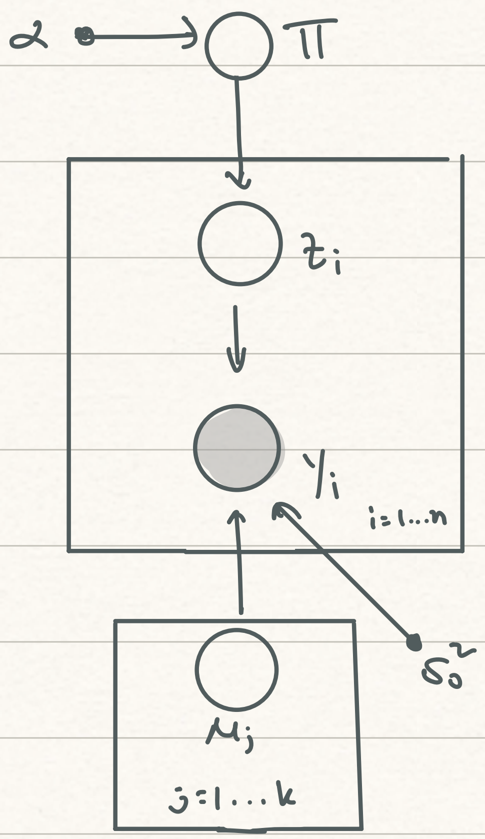
mean concentration



Draws from a 3-dimensional Dirichlet with a large concentration (left) and small concentration parameter (right). Both have the same mean.

Dirichlet-multinomial conjugacy

New graphical model:



$$p(\pi \mid -) = p(\pi \mid z, \alpha)$$

$$\propto \mathrm{Dir}(\pi; \alpha) \prod_{i=1}^{n} \mathrm{Multinoulli}(f_i; \pi)$$

$$\propto \prod_{j=1}^{k} \pi_j^{\alpha_j - 1} \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_j^{f_{ij}}$$

$$\propto \prod_{j=1}^{k} \pi_j^{\sum_i f_{ij} + \alpha_j - 1}$$

$$\propto \mathrm{Dir}(\pi; \alpha_n)$$

$$\alpha_n = [\alpha_1 + \sum_i f_{i1} \dots \alpha_K + \sum_i f_{iK}]$$

The Dirichlet is the K-dimensional generalization of the Beta distribution (K=2), and Dirichlet-multinomial conjugacy generalizes Beta-binomial conjugacy.

Bayesian exponential family mixture models (preview)

The posterior calculations above were convenient because of conditional conjugate priors. We can generalize the family of similarly tractable mixture models to a much larger set where the class-conditional likelihood is an exponential family, and the prior over the class parameters is its conjugate prior. All exponential families have a conjugate prior, a statement we will review next.

Review of exponential families

$$p(y \mid \eta) = h(y) \exp(\eta^T t(y) - a(\eta))$$

$$\propto_y h(y) \exp(\eta^T t(y))$$ “kernel”

$$a(\eta) = \log \int h(y) \exp(\eta^T t(y))\, dy$$

$\eta \in \mathbb{R}^d$ “natural parameter”

$t(y) \in \mathbb{R}^d$ “sufficient statistic”

$h(y) \in \mathbb{R}$ “base measure”

$a(\eta) \in \mathbb{R}$ “log normalizer”

e.g., the Gaussian in exponential family form

$$\mathcal{N}(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2} y - \frac{1}{2\sigma^2} y^2 - \frac{1}{2\sigma^2}\mu^2 - \log \sigma\right)$$

“link function”

$$\eta = \eta(\mu, \sigma^2) = \left[\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}\right]$$

$$t(y) = \left[y, -\tfrac{1}{2} y^2\right]$$

$$a(\eta) = \frac{1}{2}\frac{\mu}{\sigma^2} + \log \sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$$

$$h(y) = \frac{1}{2\sqrt{\pi}}$$

Conjugacy and exponential families

$$g(y_i \mid \eta) = h_\ell(y_i) \exp\left(\eta^T t_\ell(y_i) - a_\ell(\eta)\right)$$ likelihood

$$f(\eta \mid \lambda) = h_c(\eta) \exp\left(\lambda^T t_c(\eta) - a_c(\lambda)\right)$$ conjugate prior

$$\lambda = [\lambda_1, \lambda_2] \in \mathbb{R}^{\dim(\eta)+1}$$

$$\lambda_1 \in \mathbb{R}^{\dim(\eta)}$$
$$\lambda_2 \in \mathbb{R}$$

$$t_c(\eta) = [\eta, -a_\ell(\eta)]$$

$$= h_c(\eta) \exp\left(\lambda_1^T \eta - \lambda_2 a_\ell(\eta) - a_c(\lambda)\right)$$

$$\propto h_c(\eta) \exp\left(\lambda_1^T \eta - \lambda_2 a_\ell(\eta)\right)$$ kernel of conjugate prior

$$P(\eta \mid y_1 \cdots y_n, \lambda) \propto_\eta f(\eta \mid \lambda) \prod_{i=1}^{n} g(y_i \mid \eta)$$

$$\propto_\eta h_c(\eta) \exp\left(\lambda_1^T \eta - \lambda_2 a_\ell(\eta)\right) \prod_i \exp\left(\eta^T t_\ell(y_i) - a_\ell(\eta)\right)$$

$$\propto_\eta h_c(\eta) \exp\left(\eta^T \underbrace{\left[\lambda_1 + \sum_{i=1}^{n} t_\ell(y_i)\right]}_{\equiv \lambda_{n1}} - a_\ell(\eta) \underbrace{\left[\lambda_2 + n\right]}_{\equiv \lambda_{n2}}\right)$$

$$\propto f(\eta \mid \lambda_n), \quad \lambda_n = [\lambda_{n1}, \lambda_{n2}]$$

e.g., Gaussian-Gaussian conjugacy for known (unit) variance

$$\mathcal{N}(y_i; \mu, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(\mu y_i - \frac{1}{2} y_i^2 - \frac{1}{2}\mu^2\right)$$

$$= \frac{\exp\left(-\frac{1}{2} y_i^2\right)}{\sqrt{2\pi}} \exp\left(\mu y_i - \frac{1}{2}\mu^2\right)$$

$$t_\ell(y_i) = y_i$$

$$\eta = \mu$$

$$h_\ell(y_i) = \frac{\exp\left(-\frac{1}{2} y_i^2\right)}{\sqrt{2\pi}}$$

$$a_\ell(\eta) = \frac{1}{2}\mu^2 = \frac{1}{2}\eta^2$$

$$f(\eta \mid \lambda) \propto h(\eta) \exp\left(\lambda_1^T \eta - \lambda_2 a_\ell(\eta)\right)$$

$$\propto h(\eta) \exp\left(\lambda_1^T \eta - \frac{1}{2}\lambda_2 \eta^2\right)$$

$$t_c(\eta) = \left[\eta, -\frac{1}{2}\eta^2\right]$$

Bayesian exponential family mixture model

$\eta_j \sim f(\lambda) \qquad j = 1 \dots k$

$y_i \sim \sum_{j=1}^{k} \pi_j \, g(\eta_j) \qquad i = 1 \dots n$

$$P(\eta \mid y, z) \propto_\eta \prod_{j=1}^{k} (\eta_j \mid \lambda) \prod_{i=1}^{n} g(y_i \mid \eta_j)^{z_{ij}}$$

$$\propto \prod_{j=1}^{k} f(\eta_j \mid \lambda) \prod_{i=1}^{n} \exp\left(\eta_j^T t_\ell(y_i) - a_\ell(\eta_j)\right)^{z_{ij}}$$

$$\propto \prod_{j=1}^{k} f(\eta_j \mid \lambda) \exp\left(\eta_j^T \sum_{i=1}^{n} z_{ij} t_\ell(y_i) - a_\ell(\eta_j) \sum_{i=1}^{n} z_{ij}\right)$$

$$\propto \prod_{j=1}^{k} h_c(\eta_j) \exp\left(\eta_j^T \underbrace{\left[\lambda_1 + \sum_{i=1}^{n} z_{ij} t_\ell(y_i)\right]}_{\equiv \lambda_{j1}} - a_\ell(\eta_j) \underbrace{\left[\lambda_2 + \sum_{i=1}^{n} z_{ij}\right]}_{\lambda_{j2}}\right)$$

$$\propto f(\eta \mid \lambda_j)$$