22: Finite Mixture Models

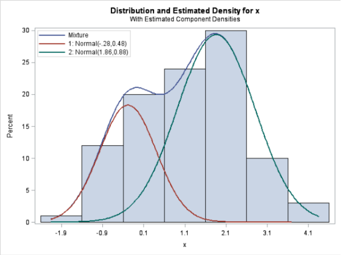
Taylor

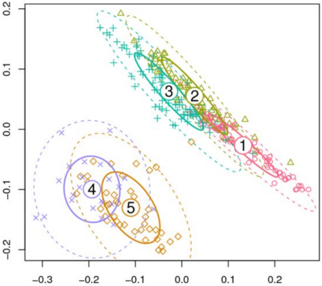
University of Virginia

Introduction

We'll take a look at **finite mixture models** now, and see how they're useful for mixture modeling.

Introduction





Notation

1. H is the number of mixtures ($h = 1, \ldots, H$)
2. $\theta = (\theta_1, \ldots, \theta_H)$ parameters for each mixture
3. $z_i = (z_{i1}, \ldots, z_{ih})$ missing data aka indicator/one-hot vector
4. $\lambda = (\lambda_1, \ldots, \lambda_H)$ parameter for $p(z_i \mid \lambda)$

and

1. $p(z_i \mid \lambda)$ distribution over missing data
2. $f(y_i \mid \theta_h)$ mixture-specific densities
3. $p(y_i \mid z_i, \theta) = \prod_{h=1}^{H} f(y_i \mid \theta_h)^{z_{ih}}$

Notation

Typically

$$p(z_i \mid \lambda) = \prod_{h=1}^{H} \lambda_h^{z_{ih}}$$

(for example $z_i = [z_{i1}, \ldots, z_{ih}] = [0, \ldots, 1, \ldots, 0]$) and

$$\begin{aligned} p(y_i \mid z_i, \theta) &= \sum_{i=1}^{H} 1_{z_{ih}=1} f(y_i \mid \theta_h) \\ &= \prod_{h=1}^{H} f(y_i \mid \theta_h)^{z_{ih}} \end{aligned}$$

so

$$p(y_i, z_i \mid \theta, \lambda) = p(y_i \mid z_i, \theta) p(z_i \mid \lambda) = \prod_{h=1}^{H} \lambda_h^{z_{ih}} f(y_i \mid \theta_h)^{z_{ih}}$$

Identifiability and Label-switching

The observed data likelihood isn't identifiable because

$$
\begin{aligned}
p(y_i \mid \theta, \lambda) &= \sum_{z_i} p(y_i \mid z_i, \theta) p(z_i \mid \lambda) \\
&= \sum_{z_i} \prod_{h=1}^{H} \lambda_h^{z_{ih}} f(y_i \mid \theta_h)^{z_{ih}} \\
&= \sum_{h} \lambda_h f(y_i \mid \theta_h) \\
&= \sum_{h} \lambda'_h f(y_i \mid \theta'_h) \\
&= p(y_i \mid \theta', \lambda')
\end{aligned}
$$

where θ' and λ' are just permuted versions of θ and λ respectively.

Identifiability and Label-switching

Watch out for exchangeable priors!

If the prior is exchangeable and the likelihood is not identifiable, then the posterior will be exchangeable:

$$\begin{aligned} p(\theta,\lambda)p(y \mid \theta,\lambda) &= p(\theta,\lambda)p(y \mid \theta',\lambda') && \text{(label switching)} \\ &= p(\theta',\lambda')p(y \mid \theta',\lambda') && \text{(exchangeable prior)} \end{aligned}$$

Identifiability and Label-switching

Watch out for exchangeable priors!

If the prior is exchangeable and the likelihood is not identifiable, then the posterior will be exchangeable:

$$\begin{aligned} p(\theta,\lambda)p(y \mid \theta,\lambda) &= p(\theta,\lambda)p(y \mid \theta',\lambda') && \text{(label switching)} \\ &= p(\theta',\lambda')p(y \mid \theta',\lambda') && \text{(exchangeable prior)} \end{aligned}$$

This means that there is no information about mixture-specific parameters.

Gibbs sampling

In a Gibbs sampling algorithm, we alternate between sampling from these conditionals:

1. $p(z \mid y, \theta, \lambda)$
2. $p(\theta, \lambda \mid z, y)$

where $y = (y_1, \ldots, y_n)$ and $z = (z_1, \ldots, z_n)$ (an $n \times h$ matrix)

Gibbs sampling

$$
\begin{aligned}
p(z \mid y, \theta, \lambda) &\propto p(\theta, \lambda) \prod_{i=1}^{n} p(y_i \mid z_i, \theta) p(z_i \mid \lambda) \\
&\propto \prod_{i=1}^{n} p(y_i \mid z_i, \theta) p(z_i \mid \lambda) \\
&= \prod_{i=1}^{n} \prod_{h=1}^{H} [\lambda_h f(y_i \mid \theta_h)]^{z_{ih}}
\end{aligned}
$$

So each z_i is Multinomial with probabilities proportional to

$$[\lambda_1 f(y_i \mid \theta_1)], \ldots, [\lambda_H f(y_i \mid \theta_H)]$$

Gibbs sampling

For the other conditional posterior:

$$p(\theta, \lambda \mid z, y) \propto p(\theta, \lambda) p(y \mid z, \theta) p(z \mid \lambda)$$

Note if $p(\theta, \lambda) = p(\theta)p(\lambda)$, then the posterior factors, too.

You can't really say any more without more details on the model.

Gibbs sampling: Example

Instead of a one-hot representation, we'll use $z_i \in \{1, \ldots, H\}$.

Gibbs sampling: Example

Instead of a one-hot representation, we'll use $z_i \in \{1, \ldots, H\}$.

Here's the complete-data likelihood:

1. $f(y_i \mid \theta_h) = \frac{1}{\sqrt{2\pi\tau_h^2}} \exp\left[-\frac{(y_i - \mu_h)^2}{2\tau_h^2}\right]$
2. $p(y_i \mid z_i, \theta) = \prod_h [f(y_i \mid \theta_h)]^{z_{ih}}$
3. $p(z_i = h \mid \lambda) = \lambda_h$
4. $p(z_i \mid \lambda) = \prod_h \lambda_h^{z_{ih}}$

Gibbs sampling: Example

Instead of a one-hot representation, we'll use $z_i \in \{1, \ldots, H\}$.

Here's the complete-data likelihood:

1. $f(y_i \mid \theta_h) = \frac{1}{\sqrt{2\pi\tau_h^2}} \exp\left[-\frac{(y_i - \mu_h)^2}{2\tau_h^2}\right]$
2. $p(y_i \mid z_i, \theta) = \prod_h [f(y_i \mid \theta_h)]^{z_{ih}}$
3. $p(z_i = h \mid \lambda) = \lambda_h$
4. $p(z_i \mid \lambda) = \prod_h \lambda_h^{z_{ih}}$

The priors for $(\theta_1, \ldots, \theta_H) = (\mu_1, \tau_1^2, \ldots, \mu_H, \tau_H^2)$ require us to pick μ_0, κ, a_τ, and b_τ:

1. $p(\mu_h \mid \tau_h^2) = \frac{1}{\sqrt{2\pi\kappa\tau^2}} \exp\left[-\frac{(\mu_h - \mu_0)^2}{2\kappa\tau_h^2}\right]$
2. $p(\tau_h^2) = \text{Inv-Gamma}(a_\tau, b_\tau)$.

Gibbs sampling: Example

Instead of a one-hot representation, we'll use $z_i \in \{1, \ldots, H\}$.

Here's the complete-data likelihood:

1. $f(y_i \mid \theta_h) = \frac{1}{\sqrt{2\pi\tau_h^2}} \exp\left[-\frac{(y_i - \mu_h)^2}{2\tau_h^2}\right]$
2. $p(y_i \mid z_i, \theta) = \prod_h [f(y_i \mid \theta_h)]^{z_{ih}}$
3. $p(z_i = h \mid \lambda) = \lambda_h$
4. $p(z_i \mid \lambda) = \prod_h \lambda_h^{z_{ih}}$

The priors for $(\theta_1, \ldots, \theta_H) = (\mu_1, \tau_1^2, \ldots, \mu_H, \tau_H^2)$ require us to pick μ_0, κ, a_τ, and b_τ:

1. $p(\mu_h \mid \tau_h^2) = \frac{1}{\sqrt{2\pi\kappa\tau^2}} \exp\left[-\frac{(\mu_h - \mu_0)^2}{2\kappa\tau_h^2}\right]$
2. $p(\tau_h^2) = \text{Inv-Gamma}(a_\tau, b_\tau)$.

Last,

1. $p(\lambda_1, \ldots, \lambda_H) \propto \prod_{h=1}^{H} \lambda^{a_h - 1}$

Gibbs sampling: Example

Overview: we derive the following two distributions

1. $p(z \mid y, \theta, \lambda)$
2. $p(\theta, \lambda \mid z, y) = p(\theta \mid z, y) p(\lambda \mid z, y)$.

The second distribution factors by the reasoning we used in slide 10.

Gibbs sampling: Example

Continuing on now with specific distributions...

$$
\begin{aligned}
p(z \mid y, \theta, \lambda) &\propto \prod_{i=1}^{n} \prod_{h=1}^{H} [\lambda_h f(y_i \mid \theta_h)]^{z_{ih}} \\
&= \prod_{i=1}^{n} \prod_{h=1}^{H} \left[\lambda_h \frac{1}{\sqrt{2\pi\tau_h^2}} \exp\left[-\frac{(y_i - \mu_h)^2}{2\tau_h^2} \right] \right]^{z_{ih}}
\end{aligned}
$$

Programming this will be easier, though, if you use `dnorm` and `rmultinom`.

Gibbs sampling: Example

Continuing on now with specific distributions...

$$
\begin{aligned}
p(\lambda \mid z, y) &\propto p(\theta)p(\lambda)p(y \mid z, \theta)p(z \mid \lambda) \\
&\propto p(\lambda)p(z \mid \lambda) \\
&\propto \left[\prod_{h=1}^{H} \lambda^{a_h - 1}\right] \left[\prod_{i=1}^{n} \prod_{h=1}^{H} \lambda_h^{z_{ih}}\right] \\
&= \prod_{h=1}^{H} \lambda^{a_h + n_h - 1}
\end{aligned}
$$

where $n_h = \sum_{i=1}^{n} 1_{z_i = h}$

Gibbs sampling: Example

Continuing on now with specific distributions...

$$
\begin{aligned}
p(\theta \mid z, y) &\propto p(\theta)p(\lambda)p(y \mid z, \theta)p(z \mid \lambda) \\
&\propto p(\theta)p(y \mid z, \theta) \\
&\propto p(\mu, \tau^2)p(y \mid z, \mu, \tau^2)
\end{aligned}
$$

where $\mu = (\mu_1, \ldots, \mu_H)$, $\tau^2 = (\tau_1^2, \ldots, \tau_H^2)$, and $n_h = \sum_{i=1}^n 1_{z_i=h}$.

Gibbs sampling: Example

The "Normal" part of the Normal-Inverse-Gamma:

$$\begin{aligned} p(\mu, \tau^2 \mid z, y) &\propto p(\mu, \tau^2) p(y \mid z, \mu, \tau^2) \\ &= \left[\prod_{h=1}^{H} p(\mu_h \mid \tau_h^2) p(\tau_h^2) \right] \left[\prod_{i=1}^{n} \prod_{h=1}^{H} f(y_i \mid \mu_h, \tau_h^2)^{z_{ih}} \right]. \end{aligned}$$

Gibbs sampling: Example

The "Normal" part of the Normal-Inverse-Gamma:

$$
\begin{aligned}
p(\mu, \tau^2 \mid z, y) &\propto p(\mu, \tau^2) p(y \mid z, \mu, \tau^2) \\
&= \left[\prod_{h=1}^{H} p(\mu_h \mid \tau_h^2) p(\tau_h^2) \right] \left[\prod_{i=1}^{n} \prod_{h=1}^{H} f(y_i \mid \mu_h, \tau_h^2)^{z_{ih}} \right].
\end{aligned}
$$

For each h

$$
p(\mu_h, \tau_h^2) p(y \mid z, \mu_h, \tau_h^2) = p(\mu_h \mid \tau_h^2) p(\tau_h^2) \prod_{i=1}^{n} f(y_i \mid \mu_h, \tau_h^2)^{z_{ih}}.
$$

will be a Normal-Inverse-Gamma distribution.

The "Normal" part of the Normal-Inverse-Gamma (continued)

$$
\begin{aligned}
& p(\mu_h \mid \tau_h^2, y, z) \\
& \propto p(\mu_h \mid \tau_h^2) \prod_{i=1}^{n} f(y_i \mid \mu_h, \tau_h^2)^{z_{ih}} \\
& \propto \frac{1}{\sqrt{2\pi\kappa\tau^2}} \exp\left[-\frac{(\mu_h - \mu_0)^2}{2\kappa\tau_h^2}\right] \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\tau_h^2}} \exp\left[-\frac{(y_i - \mu_h)^2}{2\tau_h^2}\right]\right]^{z_{ih}} \\
& \propto \exp\left[-\frac{1}{2}\left\{\frac{(\mu_h - \mu_0)^2}{\kappa\tau_h^2} + \frac{\sum_{i:z_i=h}(y_i - \mu_h)^2}{\tau_h^2}\right\}\right]
\end{aligned}
$$

For more info see page 534.

Gibbs sampling: Example

The "Inverse-Gamma" part of the Normal-Inverse-Gamma

$$
\begin{aligned}
&p(\tau_h^2 \mid y, z) \\
&\propto p(\mu_h \mid \tau_h^2) p(\tau_h^2) \prod_{i=1}^{n} f(y_i \mid \mu_h, \tau_h^2)^{z_{ih}} \\
&\propto \frac{1}{\sqrt{2\pi\kappa\tau^2}} \exp\left[-\frac{(\mu_h - \mu_0)^2}{2\kappa\tau_h^2}\right] (\tau^2)^{-(a_\tau + 1)} \exp\left[-\frac{b_\tau}{\tau_h^2}\right] \times \\
&\qquad \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\tau_h^2}} \exp\left[-\frac{(y_i - \mu_h)^2}{2\tau_h^2}\right]\right]^{z_{ih}} \\
&\propto \exp\left[-\left\{b_\tau + \frac{(\mu_h - \mu_0)^2}{2\kappa} + \frac{\sum_{i:z_i=h}(y_i - \mu_h)^2}{2}\right\} \frac{1}{\tau_h^2}\right] (\tau^2)^{-\left(\frac{n_h}{2} + \alpha_\tau + 1\right) - 1/2}
\end{aligned}
$$