

Mixture Models

Mark Andrews & Thom Baguley

May 12, 2016

Mixture models

- ▶ Given a set of observed variables $y_1, y_2 \dots y_n$, it is common to model them as distributed according to a parametric density function $P(y|\theta)$ whose parameter θ are unknown.
- ▶ For example, we could model $y_1, y_2 \dots y_n$ as

$$y_i \sim N(\mu, \sigma^2), \quad \text{for } i \in 1 \dots n.$$

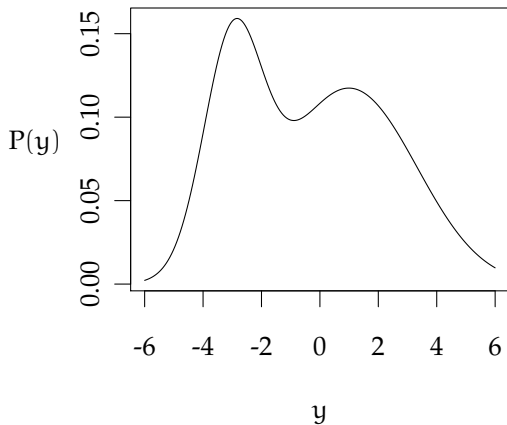
where μ and σ are unknown.

- ▶ We could, however, model $y_1, y_2 \dots y_n$ as distributed according to a finite *mixture* of parametric density functions. For example,

$$y_i \sim \sum_{k=1}^K N(\mu_k, \sigma_k) P(\mu_k, \sigma_k), \quad \text{for } i \in 1 \dots n.$$

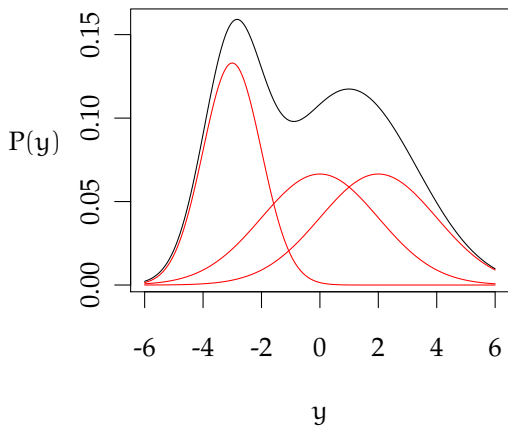
Mixture Models

A mixture of Gaussians density function.



Mixture Models

A mixture of Gaussians density function with component distributions shown in red.



Mixture models as latent variable models

- If $y_1, y_2 \dots y_n$ are modelled as

$$y_i \sim \sum_{k=1}^K N(\mu_k, \sigma_k) P(\mu_k, \sigma_k), \quad \text{for } i \in 1 \dots n,$$

this is equivalent to each y_i corresponding to a discrete latent variable $x_i \in \{1, 2 \dots K\}$ and

$$y_i \sim N(\mu_{[x_i]}, \sigma_{[x_i]}), \quad \text{for } i \in 1 \dots n.$$

and

$$x_i \sim P(\pi_k), \quad \text{where } \pi_k = P(\mu_k, \sigma_k).$$

Mixture models as latent variable models

- ▶ A general way of writing a mixture model is

$$P(y_i) \sim \sum_{k=1}^K P(y|\theta, x_i = k)P(x_i = k|\pi), \quad \text{for } i \in 1 \dots n.$$

- ▶ This is equivalent to the following generative model:
For $i \in \{1, 2 \dots n\}$,

$$x_i \sim \text{Categorical}(\pi),$$

$$y_i \sim P(y|\theta_{[x_i]})$$

Bayesian inference in mixture models as latent variable models

- Given the general way form

$$P(y_i) \sim \sum_{k=1}^K P(y|\theta, x_i = k)P(x_i = k|\pi), \quad \text{for } i \in 1 \dots n.$$

the observed data are

$$D = y_1, y_2 \dots y_n,$$

while the parameters

$$\theta = \theta_1, \theta_2 \dots \theta_K$$

and the values of the latent variables

$$x_1, x_2 \dots x_n$$

are unknown.

Bayesian inference in mixture models as latent variable models

- ▶ The full posterior is

$$P(\theta, \pi, x_{1:n} | y_{1:n}) \propto P(\pi)P(\theta) \prod_{i=1}^n P(y_i | \theta, x_i = k)P(x_i = k | \pi)$$

- ▶ We can often sample from this using a Gibbs sampler. For example, the conditional distributions

$$P(x_i = k | \theta, \pi, y_{1:n}) \propto P(y_i | x_i = k, \theta)P(x_i = k | \pi)$$

and

$$P(\theta_k | x_{1:n}, y_{1:n}, \pi) \propto P(\theta_k) \prod_{\{i: x_i = k\}} P(y_i | \theta_k)$$

and

$$P(\pi | x_{1:n}) \propto P(x_{1:n} | \pi)P(\pi)$$

are often analytically tractable.

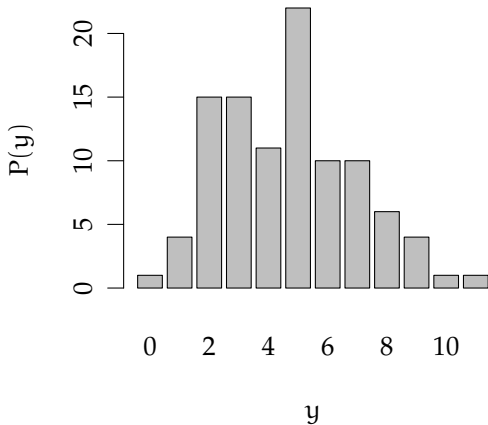
Example: Zero-Inflated Poisson distribution

- ▶ A zero inflated Poisson model is a simple two component mixture model.
- ▶ According to this model, the observed $y_1, y_2 \dots y_n$ are assumed to generated by

$$y_i \sim \begin{cases} \text{Poisson}(\lambda) & \text{if } x_i = 0, \\ 0, & \text{if } x_i = 1 \end{cases},$$
$$x_i \sim \text{Bernoulli}(\pi).$$

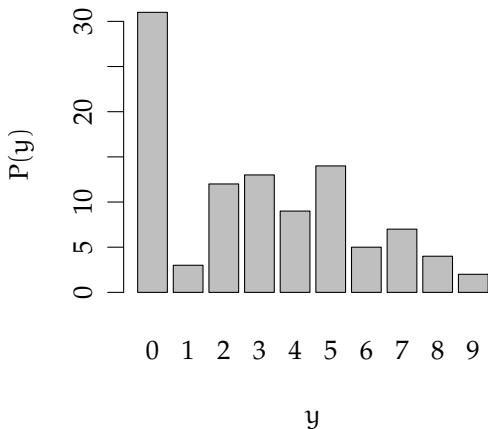
Poisson Distribution

A sample from a Poisson distribution with $\lambda = 5$.



Zero-Inflated Poisson Distribution

A sample from a Zero-inflated Poisson model, with Poisson distribution component with $\lambda = 5$, and probability of zero-model being .25.



Dirichlet Process mixture models

- ▶ Dirichlet Process mixture model can overcome the problem of choosing K , the number of component distributions.
- ▶ In a Dirichlet Process mixture model, our generative model is
For $i \in \{1, 2 \dots n\}$,

$$\theta_i \sim \text{DP}(\alpha G_0),$$

$$y_i \sim P(y|\theta_i)$$

which is equivalent to

$$P(y_i) \sim \sum_{k=1}^{\infty} P(y|\theta, x_i = k)P(x_i = k|\pi), \quad \text{for } i \in 1 \dots n.$$

Dirichlet Process mixture models

- ▶ A Dirichlet Process with *base measure* G_0 and *concentration* parameter α can be defined by following density:

$$f(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

where each π_k is drawn from a *stick-breaking* prior with parameter α , each θ_k is drawn G_0 , and

$$\delta_{\theta_k}(\theta)$$

is an function over θ that takes the value of 1 if $\theta = \theta_k$ and 0 otherwise.

- ▶ In other words, a Dirichlet Process $f(\theta)$ is an infinite distribution of *spikes* on the θ parameter space.