Mixture Models

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

04/16/2024

Finite mixture models

- Let $f_1, \ldots, f_K$ be densities with weights $q_1, \ldots, q_k$ where

$$q_1 + q_2 + \cdots + q_K = 1 \text{ and } q_k \geq 0 \, \forall k.]$$

- Assume $y_1, \ldots, y_n$ are iid with density

$$p(y_i) = \sum_{k=1}^{K} q_k f_k(y_i).$$

- Equivalently,

$$y_i \sim \begin{cases} f_1 \text{ with probability } q_1 \\ \vdots \\ f_K \text{ with probability } q_k \end{cases}$$

- Useful for modeling complex distributions.
- *Mixture* of simpler *component* distributions f_k.

Dirichlet distribution

- In a Bayesian framework, put a prior on $q = (q_1, \ldots, q_K)$
- A *Dirichlet* prior is commonly used for q
  - Parameterized by *concentration* values $\alpha = (\alpha_1, \ldots, \alpha_K)$:

$$\pi(q) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} q_k^{\alpha_k - 1}$$

    for $q_1 + q_2 + \cdots + q_K = 1$ and $q_k \geq 0 \; \forall \; k$.
  - $B(\cdot)$ is the multivariate beta function, defined by the gamma function Γ:

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

- For $K = 2$, the Dirichlet(α_1, α_2) = Beta(α_1, α_2)

Dirichlet distribution

- The domain of the Dirichlet is the $K-1$ unit simplex

$$\triangle^{K-1} = \left\{ q_1, \ldots, q_K \in \mathbb{R}^K : \sum_{k=1}^{K} q_k = 1 \text{ and } q_k \geq 0 \,\forall\, k \right\}$$

- $\alpha_1 = \ldots = \alpha_K = 1$ implies a uniform distribution on $\triangle^{K-1}$
- Dirichlet-multinomial model:
  - Let $z_1, \ldots, z_n$ be iid variables from K categories
  - $z_i \in \{1, \ldots, K\}$ with $p(z_i = k) = q_k$
  - Let n_k be the number of $z_i's$ from category k.

$$(n_1, \ldots, n_k) \sim \text{Multinomial}(n, q)$$

  - The Dirichlet(α) is conjugate for q:

$$p(q \mid \mathbf{z}) = \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_1 + n_K)$$

Bayesian finite mixture model

- We can also put a prior on $(f_1, \ldots, f_K)$
- Commonly assume f_k's come from same parametric family:

$$f_k(\cdot) = f(\cdot \mid \theta_k)$$

and put a prior on $\theta_1, \ldots, \theta_k$

- For example, $f_k = \text{Normal}(\mu_k, \sigma_k^2)$ with independent priors

$$\mu_k \sim \text{Normal}(\mu_0, \tau^2)$$

$$\sigma_k^2 \sim IG(a, b)$$

on $\theta_k = (\mu_k, \sigma_k^2)$ for $k = 1, \ldots, K$.

Estimation

- Let $z_i \in \{1, \ldots, K\}$ indicate the component that generated y_i
- Gibbs sample from full conditionals for $\mathbf{z}$, q, and $(\theta_1, \ldots, \theta_k)$:
  - Draw z_i, for $i = 1, \ldots, n$, by

$$p(z_i = k \mid \mathbf{y}, q, \theta) \propto q_k f(y_i \mid \theta_k)$$

  - Draw q by

$$p(q \mid \mathbf{y}, \theta, \mathbf{z}) = \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$$

    where $n_k = \sum_{i=1}^{n} \mathbb{1}_{\{z_i = k\}}$
  - Draw θ_k, for $k = 1, \ldots, K$, from

$$p(\theta_k \mid \mathbf{y}, q, z_i) \propto \pi(\theta_k) f(\mathbf{y}_k \mid \theta_k)$$

    where $\mathbf{y}_k = \{y_i : z_i = k\}$. This may require a Metropolis step, if conjugate prior s are not used.

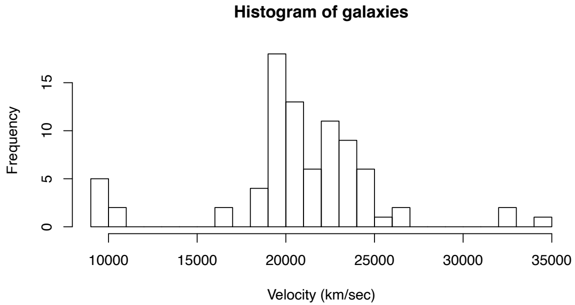
Example: Galaxies

- Measurements taken from 82 galaxies in Corona Borealis region of space [1]
- Consider velocity in km/sec for each galaxy
    - Galaxies with a higher velocity are farther from earth.

[1] Roeder, K. Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *JASA* 85, 617-624.

Example: Galaxies

- The distribution of velocities ($\mathbf{y}$) is very multimodal



http://www.ericfrazerlock.com/Mixture_Models_Rcode1.r

Example: Galaxies

- Model $\mathbf{y}$ as a mixture of normal densities:

$$p(y_i \mid q, (\mu_k, \sigma_k^2)_{k=1}^K) \stackrel{iid}{=} \sum_{k=1}^{K} q_k \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

for $i = 1, \ldots, n$.

- Use a uniform Dirichlet prior for q

$$q \sim \text{Dirichlet}(\mathbf{1})$$

- Use a mildy informative normal-inverse-gamma prior for μ_k, σ_k^2:

$$\mu_k \stackrel{iid}{\sim} \text{Normal}(20000, 10^9)$$

$$\sigma_k^2 \stackrel{iid}{\sim} IG(2, 10^8)$$

Example: Galaxies

- Estimate for $K = 4$ clusters
- For Gibbs sampling, initialize:
  - Draw μ_k, σ_k^2 from prior for $k = 1, \ldots, 4$
  - Set $q_k = 1/4$ for $k = 1, \ldots, 4$
- The full conditional for each μ_k σ_k^2 can be drawn from a normal-inverse-gamma posterior.
- Run Gibbs sampling for $T = 50000$ iterations

R CODE

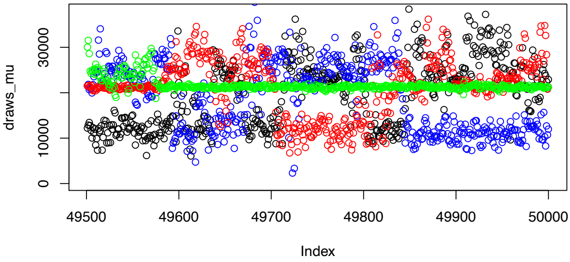
```
for(t in 1:T){ ##Run gibbs sampler
 ###Generate component indicators Z
 for(k in 1:K) probs[k,] = q[k]*dnorm(y,mu[k],sqrt(sigma2[k]))
 for(i in 1:n) Z[i] = which(rmultinom(1,1,probs[,i])==1)

 ###Generate qs
 nk = c()
 for(k in 1:K) nk[k] = sum(Z==k)
 q = rdirichlet(1,alpha+nk)

 ###generate mu_k, sigma2_k from normal-inverse-gamma posterior
 for(k in 1:K){
   a_post = a+nk[k]/2-1/2
   b_post = b+(1/2)*sum((y[Z==k]-mean(y[Z==k]))^2)
   sigma2[k] = 1/rgamma(1,a_post,b_post)
   post_mu_mean = (sigma2[k]*mu0+tau2*sum(y[Z==k]))/(nk[k]*tau2+
   post_mu_var = sigma2[k]*tau2/(nk[k]*tau2+sigma2[k])
   mu[k] = rnorm(1,post_mu_mean,sqrt(post_mu_var))
 }
```

R CODE (continued)

```
 ##Store draws
 draws_q[t,] = q
 draws_mu[t,] = mu
 draws_sigma2[t,] = sigma2
 ###Compute density over grid
 x = seq(from = 5000, to=40000, length.out=200)
 Dens = rep(0,200)
 for(k in 1:K) Dens = Dens+q[k]*dnorm(x,mu[k],sqrt(sigma2[k
 draws_dens[t,] = Dens
}
```

Example: Galaxies

- Plot of posterior draws for $\mu_1, \mu_2, \mu_3, \mu_4$, last 500 iterations:



http://www.ericfrazerlock.com/Mixture_Models_Rcode1.r

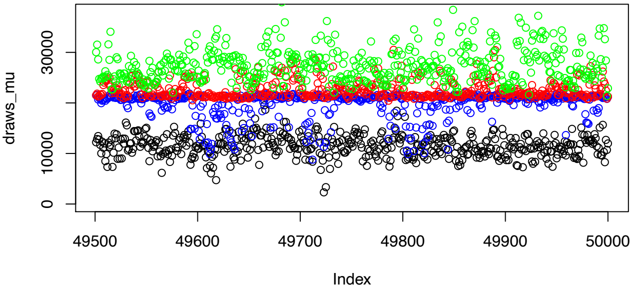
- Chains cross – "label switching"

R CODE

```
for(t in 1:T){
  Order = order(draws_mu[t,])
  draws_q[t,] = draws_q[t,Order]
  draws_mu[t,] = draws_mu[t,Order]
  draws_sigma2[t,] = draws_sigma2[t,Order]
}
```

Example: Galaxies

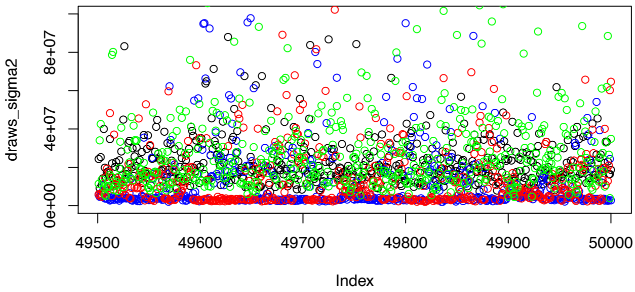
- Maintaining order $\mu_1 < \mu_2 < \mu_4 < \mu_4$:



- Keep these labels for subsequent computation

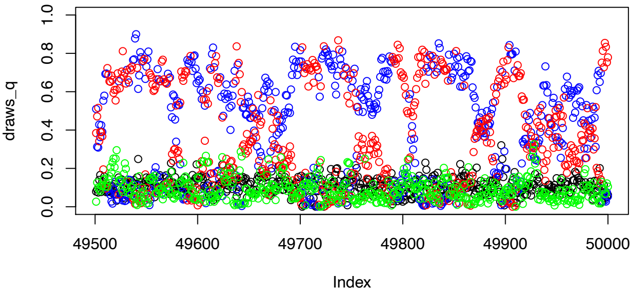
Example: Galaxies

- Final 500 draws for $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$:



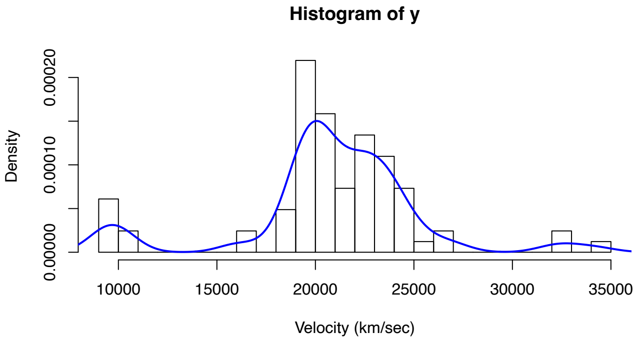
Example: Galaxies

- Final 500 draws for q_1, q_2, q_3, q_4:



Example: Galaxies

- Density estimate based on posterior draws:



Example: Galaxies

- Component estimates:
  - $\mu_1 = 11529$, $\sigma_1^2 = 28397532$, $q_1 = 0.10$
  - $\mu_2 = 19501$, $\sigma_2^2 = 24032751$, $q_2 = 0.40$
  - $\mu_3 = 22566$, $\sigma_3^2 = 18692368$, $q_3 = 0.40$
  - $\mu_4 = 27818$, $\sigma_4^2 = 34797849$, $q_4 = 0.10$

Choosing K

- Number of components (i.e., *clusters*) K can be selected using standard model selection tools
- DIC, BIC, cross-validation, etc.
- Alternatively, put a prior on K
- Or, choose K large and use small values for α
  - Some components will have negligable probability
- Letting $K \to \infty$ with $\alpha_k = c/K$ for all k gives a *Dirichlet process*.