A Short Course on Bayesian Nonparametrics

Lecture 2 - Introduction to Dirichlet process mixture models

Abel Rodriguez - UC, Santa cruz

Universidade Federal Do Rio de Janeiro
March, 2011

Two-component mixtures

A two-component mixture model

$$y_i|\omega, \{\mu_k\}_{k=1}^2, \{\sigma_k^2\}_{k=1}^2 \sim_{iid} \omega \mathrm{N}(y_i|\mu_1, \sigma_1^2) + (1-\omega)\mathrm{N}(y_i|\mu_2, \sigma_2^2)$$

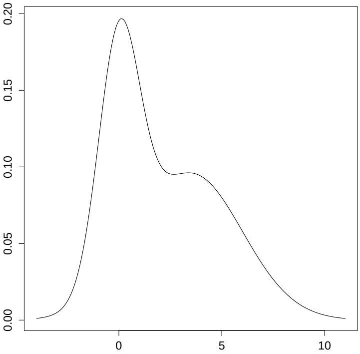
Another way to write the likelihood:

$$y_i \sim_{iid} \mathrm{N}(\mu_{\xi_i}, \sigma_{\xi_i}^2)$$

and

$$\Pr(\xi_i = 1) = \omega = 1 - \Pr(\xi_i = 2)$$

also iid.



General finite mixture

More generally

$$y_i|\{\omega_k\}_{k=1}^K, \{\vartheta_k\}_{k=1}^K \sim_{iid} \sum_{k=1}^K \omega_k \psi(y_i|\vartheta_k) \quad \sum_{k=1}^K \omega_k = 1 \quad \vartheta_k \in \Theta$$

A typical set of priors for this model is

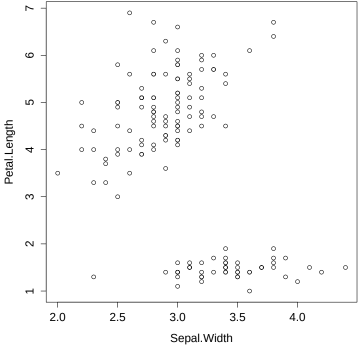
$$(\omega_1, \ldots, \omega_k) \sim \mathrm{Dir}(a_1, \ldots, a_k) \qquad \vartheta_k \sim_{iid} H$$

Posterior inference:

$$(\omega_1, \ldots, \omega_k)|\{\xi_i\}, y \qquad \xi_i|\{\theta_k\}, \{\omega_k\}, y \qquad \theta_k|\{\xi_i\}, y$$

Applications: Model-based clustering

- Divide observations into "homogeneous" $\Rightarrow \{\xi_i\}$ unknown for all observations.
- What homogeneous means depends on what kernel ψ is used $\Rightarrow$ Under multivariate normality, Bayesian K-means clustering.
- Two clear clusters in the Iris dataset (truth is three)



Clasification

- Given examples belonging to different classes (training set), allocate new observations (test set).
- Classification problems can be framed in a similar way $\Rightarrow$ In this case, we know what the value of ξ_i for i in the training set, and we want to infer ξ_i for i in the test set.
- Under multivariate normality, Bayesian linear and quadratic discriminant analysis (LDA, QDA) can be recovered.
- Check Fraley & Raftery (2002)

Density estimation

- Density estimation problems ⇒ With enough components, a mixture of normals can approximate any continuous distribution.
- Can be considered a Bayesian version of kernel density estimation (KDE).
- Some useful representation theorems
  - Location-scale mixtures of normals can approximate arbitrarily well any density on the real line (Lo, 1984; Ferguson, 1983; Escobar and West, 1995). Analogously, for densities on R^d (West et al., 1994; Müller et al., 1996).
  - Unimodal symmetric densities on the real line can be represented as mixtures of uniform distributions (Brunner and Lo, 1989; Brunner, 1995; Lavine and Mockus, 1995; Kottas and Gelfand, 2001).

Fitting a finite mixture of Gausians

$$y_i \sim \mathrm{N}(\theta_{\xi_i}, \sigma^2) \quad \Pr(\xi_i = k) = \omega_k \quad \theta_k \sim \mathrm{N}(\theta_0, D_0) \quad \omega \sim \mathrm{Dir}(a)$$

```
library(MCMCpack)

sample.omega <- function(y, xi, a){
KK <- length(a)
omega <- rdirichlet(1, a + tabulate(xi, nbins
= KK))
return(omega)
}

sample.xi <- function(y, theta, sigma2,
omega){
NN <- length(y)
xi <- rep(0, NN)
for(i in 1:NN){
qq <- log(omega) + dnorm(y[i], theta,
sqrt(sigma2), log=T)
qq <- exp(qq - max(qq))
qq <- qq/sum(qq)
xi[i] <- sample(1:KK, 1, T, qq)
}
return(xi)
}
```

```
sample.theta <- function(y, KK, xi, sigma2,
theta0, D0){
theta <- rep(0, KK)
for(k in 1:KK){
nk <- sum(xi==k)
sk <- sum(y[xi==k])
theta[k] <- rnorm(1, (sx/sigma2 +
theta0/D0)/(nx/sigma2 + 1/D0) ,
sqrt(1/(nx/sigma2 + 1/D0)) )
}
return(theta)
}
```

Homework

1. Generate a sample $n = 30$ observations from the mixture

$$y_i \sim 0.4\mathrm{N}(y_i|0,1) + 0.6\mathrm{N}(y_i|3.5, 2.5^2)$$

2. Modify the code above to fit a finite mixture models with

$$\psi(y_i|\theta_{\xi_k}, \sigma^2_{\xi_k}) = \mathrm{N}(y_i|\theta_{\xi_i}, \sigma^2_{\xi_i})$$

   Take $\omega \sim \mathrm{Dir}(1, \dots, 1)$ and let H be a normal inverse Gamma distribution.

3. Discuss how to pick the hyperparameters for H.
4. Fit the models with both $K = 2$ and $K = 50$. Compare the results and discuss.

Homework

Some issues that you might encounter:

1. In what sense should we compare the fits?
2. How to summarize the posterior distribution on $\{\xi_i\}$?
3. How do you construct an estimate for $p(y_{n+1}|y_1, \ldots, y_n)$?
4. Label switching?

Alternative formulations

- A fancier way to write the finite mixture model

$$y_i \sim \int \psi(y_i|\theta_i)dG(\theta_i) \qquad G(\cdot) = \sum_{k=1}^{K} \omega_k \delta_{\theta_k}(\cdot)$$

  where $\delta_\vartheta(\cdot)$ is a degenerate measure putting probability one on the value ϑ.

- Hence, a prior on $(\{\omega_k\}, \{\vartheta_k\})$ is equivalent to a prior on the discrete measure G.
- $G \sim \mathrm{DP}(\alpha, H) \Rightarrow$ Dirichlet process mixture model.

Dirichlet process mixtures

- The model is

$$y_i|G \sim \int \psi(y_i|\theta)dG(\theta) \qquad G \sim \mathrm{DP}(\alpha, H)$$

- Consider the DPM prior as a “smoothed version” of the DP prior (just like the KDE is a smoothed version of the ECDF).
- Can model both discrete and continuous distributions (simply by changing the kernel).

Dirichlet process mixtures

- Useful to rewrite as

$$y_i|\theta_i \sim \psi(y_i|\theta_i) \qquad \theta_i|G \sim G \qquad G \sim \mathrm{DP}(\alpha, H)$$

so that the θ_i's can be interpreted as subject-specific random effects.

- Or, by using the constructive definition of the DP,

$$y_i|\{\omega_k\}, \{\vartheta_k\} \sim \sum_{k=1}^{\infty} \omega_k \psi(y_i|\vartheta_k)$$

with $\vartheta_k \sim_{iid} H$, $\omega_k = z_k \prod_{l<k}\{1 - z_l\}$, and $z_k \sim_{iid} \mathrm{beta}(1, \alpha)$. Infinite potential number of clusters, in practice only a small number K are occupied.

Other characterizations

Limit of a finite mixture model (Ishwaran & Zarepour, 2002). If

$$p^K(y|\{\vartheta_k\}, \{\omega_k^*\}) = \sum_{k=1}^{K} \omega_k^* \psi(y|\vartheta_k)$$

$$(\omega_1^*, \ldots, \omega_K^*) \sim \text{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

$$\vartheta_k \sim H$$

then

$$\lim_{K\to\infty} p^K(y) \stackrel{D}{=} \int \psi(y|\theta) dG(\theta)$$

where $G \sim \text{DP}(\alpha, H)$

Parameter elicitation

- Note that

$$\begin{aligned} \mathrm{E}(y_i) &= \mathrm{E}_H\{\mathrm{E}(y_i|\theta_i)\} \\ \mathrm{Var}(y_i) &= \mathrm{Var}_H\{\mathrm{E}(y_i|\theta_i)\} + \mathrm{E}_H\{\mathrm{Var}(y_i|\theta_i)\} \end{aligned}$$

  Which is helpful to elicit hyperparameters associated with H.

- Also, α controls how many distinct values are in the sample $\theta_1, \ldots, \theta_n \Rightarrow K =$ Number of **occupied** clusters (Antoniak, 1974).

$$\Pr(K = m \mid \alpha) = c_n(m) n! \alpha^m \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \qquad m = 1, \ldots, n,$$

  For moderately large n

$$\mathrm{E}(K \mid \alpha) \approx \alpha \log\left(\frac{\alpha + n}{\alpha}\right)$$

Limiting cases

- If $\alpha \to 0$, this is a single-component mixture

$$y_i|\theta \sim_{iid} \psi(y_i|\theta) \qquad\qquad \theta \sim H$$

- If $\alpha \to \infty$, we have one component per observations.

$$y_i \sim_{iid} \int \psi(y_i|\theta) dH(\theta)$$

(there is nothing unknown in the distribution of y_i unless a prior on the hyperparameters of H is used).

Full hierarchical formulation

- A more useful (semi)parametric model

$$\begin{aligned} y_i|\theta_i, \phi &\sim \psi(y_i|\theta_i, \phi) \\ \theta_i|G &\sim G \\ G|\alpha, \eta &\sim \mathrm{DP}(\alpha, H_\eta) \\ \alpha, \phi, \eta &\sim p(\alpha, \phi, \eta) \end{aligned}$$

  - ϕ is a fixed effect common to all subjects (for example, it could eventually be a set of regression coefficients).
  - Letting α and η be random provides additional flexibility.
- Most often, $p(\alpha, \phi, \eta) = p(\alpha)p(\phi)p(\eta)$

Marginalized (collapsed) version of the model

- We can avoid dealing with an infinite number of parameters by integrating the unknown G.

$$
\begin{aligned}
y_i|\theta_i, \phi &\sim \psi(y_i|\theta_i, \phi) \\
(\theta_1, \ldots, \theta_n)|\alpha, \eta &\sim \mathrm{PU}(\alpha, H_\eta) \\
\alpha, \phi, \eta &\sim p(\alpha)p(\phi)p(\eta)
\end{aligned}
$$

- No need to store (sample) the ω_ks or represent the ϑ_ks that do not have observations associated with them.
- For now, we assume that ψ and H are conjugate!

Marginalized (collapsed) version of the model

- The posterior distribution is

$$p(\{\theta_i\}, \phi, \alpha, \eta | y) \propto \left\{ \prod_{i=1}^{n} \psi(y_i | \theta_i, \phi) \right\} p(\theta_1, \ldots, \theta_n | \alpha, \eta) p(\alpha) p(\phi) p(\eta)$$

- We need four sets of full conditionals:
  - $\theta_i | \theta_{-i}, \alpha, \eta, \phi, y$
  - $\alpha | \{\theta_i\}$
  - $\eta | \{\theta_i\}$
  - $\phi | \{\theta_i\}, y$
- For the moment we focus on the first one.

$$p(\theta_i | \theta_{-i}, \alpha, \eta, \phi, y) \propto \psi(y_i | \theta_i, \phi) p(\theta_1, \ldots, \theta_n | \alpha, \eta)$$

Marginalized (collapsed) version of the model

- Remember the Pòlya urn. Since θ_is are exchangeable, their prior full conditional is the same for all i.

$$p(\theta_i|\theta_{-i}, \alpha, \eta) = \sum_{k=1}^{K^{-i}} \frac{m_k^{-i}}{n-1+\alpha} \delta_{\vartheta_k^{-i}} + \frac{\alpha}{n-1+\alpha} h_\eta$$

- The negative exponent denotes quantities computed after removing the corresponding observation.
- $\vartheta_1^{-i}, \ldots, \vartheta_{K^{-i}}^{-i}$ are the unique values in θ_{-i}.
- m_k^{-i} is the size of the k-th cluster after removing observation i.
- K^{-i} is the number of clusters after eliminating observation i.
- Remember that the clusters need to be labeled continuously (requires bookkeeping when θ_i is in a cluster of its own).

Marginalized (collapsed) version of the model

- The corresponding full conditional posterior is

$$p(\theta_i|\theta_{-i},\alpha,\eta,\phi,y) \propto \psi(y_i|\theta_i,\phi)\left\{\sum_{k=1}^{K^{-i}} m_k^{-i}\delta_{\vartheta_k^{-i}}(\theta_i) + \alpha h_\eta(\theta_i)\right\}$$

$$= \sum_{k=1}^{K^{-i}} m_k^{-i}\psi(y_i|\vartheta_k^{-i},\phi)\delta_{\vartheta_k^{-i}}(\theta_i) +$$

$$\alpha p(y_i|\phi,\eta)\frac{\psi(y_i|\theta_i,\phi)h_\eta(\theta_i)}{p(y_i|\phi,\eta)}$$

  - With prob. prop. to $m_k^{-i}\psi(y_i|\vartheta_k^{-i},\phi)$ we make $\theta_i = \vartheta_k^{-i}$.
  - With prob. prop. to $\alpha p(y_i|\phi,\eta) = \int \psi(y_i|\theta_i,\phi)h_\eta(\theta_i)d\theta_i$ we open a new component and sample θ_i from the posterior associated with the prior h_η and the likelihood $\psi(y_i|\theta_i,\phi)$.

An alternative representation

- If the PU is used directly, θ_is change only when they are reallocated to new components $\Rightarrow$ Very slow mixing.
- An improved algorithm $\Rightarrow$ Introduce indicators $\xi_i \in \mathbb{N}$ and ϑ_l such that $\theta_i = \vartheta_{\xi_i}$.
- The model can be written as

$$
\begin{aligned}
y_i|\xi_i, \{\vartheta_k\}, \phi &\sim \psi(y_i|\vartheta_{\xi_i}, \phi) \\
\vartheta_k|\eta &\sim H_\eta \\
(\xi_1, \ldots, \xi_n)|\alpha &\sim \mathrm{CRP}(\alpha) \\
\alpha, \phi, \eta &\sim p(\alpha)p(\phi)p(\eta)
\end{aligned}
$$

- Then the prior full conditional urn can be written as

$$
\xi_i|\xi_{-l}, \alpha \sim \sum_{k=1}^{K^{-i}} \frac{m_k^{-i}}{n-1+\alpha}\delta_k + \frac{\alpha}{n-1+\alpha}\delta_{K^{-i}+1}
$$

Joint posterior

- Joint posterior

$$p(\{\vartheta_k\}, \{\xi_i\}, \phi, \alpha, \eta | y) \propto \left\{ \prod_{i=1}^{n} \psi(y_i | \vartheta_{\xi_i}, \phi) \right\}$$
$$\left\{ \prod_{k=1}^{\max\{\xi_i\}} h(\vartheta_k | \eta) \right\} p(\xi_1, \ldots, \xi_n | \alpha) p(\alpha) p(\phi) p(\eta)$$

- Note that inferences on the number of components K are done indirectly through inferences on $\{\xi_k\}$ (because $K = \max\{\xi_i\}$).
- For the moment, we assume conjugacy of H and ψ.

Full conditional

- As before

$$p(\xi_i|\xi_{-i},\{\vartheta_k^{-i}\},\alpha,\eta,\phi) \propto \psi(y_i|\vartheta_{\xi_i},\phi)\left\{\sum_{k=1}^{K^{-i}} m_k^{-i}\delta_k(\xi_i) + \alpha\delta_{K^{-i}+1}(\xi_i)\right\}$$

$$= \sum_{k=1}^{K^{-i}} m_k^{-i}\psi(y_i|\vartheta_k^{-i},\phi)\delta_k(\xi_i) + \alpha\psi(y_i|\vartheta_{K^{-i}+1}^{-i},\phi)\delta_{K^{-i}+1}(\xi_i)$$

- Also

$$p(\vartheta_k^{-i}|y_{-i}) \propto \begin{cases} \left\{\prod_{\{j:\xi_j=k,j\neq i\}} \psi(y_j|\vartheta_k^{-i},\phi)\right\} h_\eta(\vartheta_k^{-i}) & k \leq K^{-i} \\ h_\eta(\vartheta_k^{-i}) & k = K^{-i}+1 \end{cases}$$

- In the conjugate case, we can integrate the ϑ_k^{-i}s (Rao-Blackwellization)

Full conditionals

- Latent indicators $\{\xi_i\}$

$$\Pr(\xi_i = k | \cdots) \propto \begin{cases} m_k^{-i} p(y_i | \{y_j : \xi_j = k, j \neq i\}, \phi, \eta) & k \leq K^{-i} \\ \alpha p(y_i | \phi, \eta) & k = K^{-i} + 1 \end{cases}$$

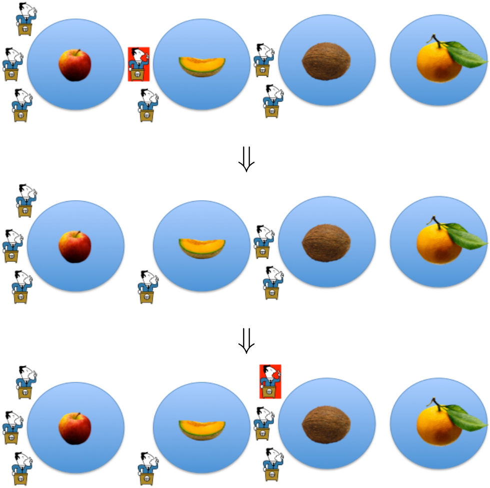
where

$$p(y_i | \{y_j : \xi_j = k, j \neq i\}, \phi, \eta) = \frac{\int \prod_{\{j : j = i \text{ or } \xi_j = k\}} \psi(y_j | \theta, \phi) dH_\eta(\theta)}{\int \prod_{\{j : \xi_j = k, j \neq i\}} \psi(y_j | \theta, \phi) dH_\eta(\theta)}$$

$$p(y_i | \phi, \eta) = \int \psi(y_i | \theta, \phi) dH_\eta(\theta)$$

The ξ_i's are **always** assumed to be labeled continuously starting at 1 (book-keeping!!!).

Collapsed Gibbs



Full conditionals

- For $\{\vartheta_k\}$

$$p(\vartheta_k | \cdots) \propto \left\{ \prod_{\{j:\xi_j=k\}} \psi(y_j|\vartheta_k, \phi) \right\} H_\eta(\vartheta_k)$$

- For ϕ

$$p(\phi | \cdots) \propto \left\{ \prod_{i=1}^{n} \psi(y_j|\vartheta_{\xi_i}, \phi) \right\} p(\phi)$$

- For η, remember that the ϑ_ks are iid samples from H_η

$$p(\eta | \cdots) \propto \left\{ \prod_{k=1}^{\max\{\xi_i\}} h(\vartheta_k|\eta) \right\} p(\eta)$$

Full conditionals

- For α, note that if $K = \max\{\xi_i\}$

$$p(\alpha | \cdots) \propto p(\alpha)\alpha^K \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = p(\alpha)\alpha^K \frac{(\alpha+n)}{\Gamma(n+1)} \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n+1)}$$
$$\propto p(\alpha)\alpha^{K-1}(\alpha+n)\int_0^1 \varsigma^{\alpha}(1-\varsigma)^{n-1} d\varsigma$$

- Hence, if $\alpha \sim \text{Gam}(a_\alpha, b_\alpha)$ then

$$\varsigma | \alpha, \cdots \sim \text{beta}(\alpha+1, n)$$
$$\alpha | \varsigma, \cdots \sim \epsilon \text{Gam}(a_\alpha + K, b_\alpha - \log\varsigma) + (1-\epsilon)\text{Gam}(a_\alpha + K - 1, b_\alpha - \log\varsigma)$$

with $\epsilon = (a_\alpha + K - 1)/(a_\alpha + K - 1 + n\{b_\alpha - \log(\varsigma)\})$.

A comparison between finite and infinite mixture samplers

Par	Inifinte mixture	Finite mixture
ξ_i	Multinomial (PU) (variable size $K^{-i}+1 \leq n$) $\Pr(\xi_i = k \mid \cdots) \propto \begin{cases} m_k^{-i}\psi(y_i \mid \vartheta_k^{-i}, \phi) & k \leq K^{-i} \\ \alpha p(y_i \mid \phi, \eta) & k = K^{-i}+1 \end{cases}$	Multinomial (fixed size K) $\Pr(\xi_i = k \mid \cdots) \propto \omega_k \psi(y_i \mid \vartheta_k, \phi)$
ϑ_k	Standard $p(\vartheta_k \mid \cdots) \propto \left\{ \prod_{\{j:\xi_j=k\}} \psi(y_j \mid \vartheta_k, \phi) \right\} h_\eta(\vartheta_k)$	Same
ω	Not needed (integrated out)	Standard $\omega \mid \cdots \sim \mathrm{Dir}_K(m_1+\alpha_1, \ldots, m_K+\alpha_K)$
ϕ	Standard $p(\phi \mid \cdots) \propto \left\{ \prod_{i=1}^n \psi(y_i \mid \vartheta_{\xi_i}, \phi) \right\} p(\phi)$	Same
α	Data augmentation	Typically not done

An example: location mixtures of normals

- $y_i|\theta_i \sim \mathrm{N}(\theta_i, \sigma^2)$, $\theta_i \sim G$, $G \sim \mathrm{DP}(\alpha, H)$ and $H = \mathrm{N}(b, B)$.
- For sampling ξ_is we need

$$p(y_i|\phi,\eta) = \int \psi(y_i|\theta_i,\sigma^2)dH(\theta_i) = \frac{1}{\sqrt{2\pi}\sqrt{B+\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i-b)^2}{B+\sigma^2}\right\}$$

and

$$p(y_i|\{y_j : \xi_j = k, j \neq i\}, \phi, \eta) = \int \psi(y_i|\theta_i,\sigma^2)p(\theta_i|\{y_j : \xi_j = k, k \neq i\})d\theta_i$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\hat{\sigma}^2_{k,-i}+\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i-\hat{\mu}_{k,-i})^2}{\hat{\sigma}^2_{k,-i}+\sigma^2}\right\}$$

with

$$\hat{\sigma}^2_{k,-i} = \left\{\frac{1}{B} + \frac{m_k^{-i}}{\sigma^2}\right\}^{-1} \quad \hat{\mu}_{k,-i} = \left\{\frac{1}{B} + \frac{m_k^{-i}}{\sigma^2}\right\}^{-1}\left\{\frac{b}{B} + \frac{1}{\sigma^2}\sum_{\{j:\xi_j=k,k\neq i\}} y_j\right\}$$

An example: location mixtures of normals

- For sampling the ϑ_ks we have $\vartheta_k | \cdots \sim \mathsf{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$ (without the $-i$).
- If $\sigma^2 \sim \mathsf{IGam}(a_\sigma, b_\sigma)$ then

$$\sigma^2 | \cdots \sim \mathsf{IGam}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2} \sum_{i=1}^{n} (y_i - \vartheta_{\xi_i})^2 \right)$$

- If $b \sim \mathsf{N}(b_0, D)$ and $B \sim \mathsf{IGam}(a_B, c_B)$

$$b | \cdots \sim \mathsf{N}\left(\left\{ \frac{1}{D} + \frac{K}{B} \right\}^{-1} \left\{ \frac{b_0}{D} + \frac{1}{B} \sum_{k=1}^{K} \vartheta_k \right\}, \left\{ \frac{1}{D} + \frac{K}{B} \right\}^{-1} \right)$$

$$B | \cdots \sim \mathsf{IGam}\left(a_B + \frac{K}{2}, c_B + \frac{1}{2} \sum_{k=1}^{K} (\vartheta_k - b)^2 \right)$$

Density estimation using DP mixture models

The predictive distribution for a new observation can be computed as

$$p(y_{n+1}|y_1\ldots,y_n) = \int \psi(y_{n+1}|\theta_{n+1},\phi)p(\theta_{n+1}|\theta_n,\ldots,\theta_{n+1},\alpha,\eta)$$
$$p(\theta_n,\ldots,\theta_1,\alpha,\eta|y_1,\ldots,y_n)d\theta_{n+1}d\theta_n\ldots d\theta_1 d\alpha d\eta$$

The B samples obtained from the MCMC allows us to construct an approximation

$$p(y_{n+1}|y_1\ldots,y_n) \approx \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{\alpha^{(b)}}{n+\alpha^{(b)}}p(y_i|\phi^{(b)},\eta^{(b)})\right.$$
$$\left. + \sum_{k}^{K^{(b)}}\frac{m_k^{(b)}}{n+\alpha^{(b)}}p(y_i|\{y_j:\xi_j=k,j\neq i\},\phi^{(b)},\eta^{(b)})\right\}$$

Implementation issues

Book-keeping for this algorithm can be tricky:

```
sample.xi <- function(y, xi, alpha, m, B){
K <- max(xi)
n <- length(xi)
for (i in 1:n){
xi.n <- xi
xi.n[i] <- 0
xi.n[-i] <- as.numeric(factor(xi[-i], labels = seq(1,length(unique(xi[-i])))))
K.n <- max(xi.n)
q <- rep(0, K.n+1)
for(k in 1:K.n) {
q[k] <- log(sum(xi.n==k)) + logpred(y[i], sum(xi.n==k), sum(y[xi.n==k]), m, B)
}
q[K.n+1] <- log(alpha) + logpred(y[i], 0, 0, m, B)
w <- exp(q - max(q))
w <- w/sum(w)
ind <- sample(1:(L.n+1), 1, replace=T, w)
xi.n[i] <- ind
xi <- xi.n
B <- B.n
L <- max(xi)
}
return(list(xi = xi, B=B))
}
```

Homework

1. Complete the code necessary to code the collapsed sampler for the location mixture of normals (If you are careful, you can reuse a good part of the code you used for the finite mixture model).
2. Compare the results you obtain here with those of the finite mixture model (take $a_\alpha = 1$ and $b_\alpha = 1$).

DPpackage

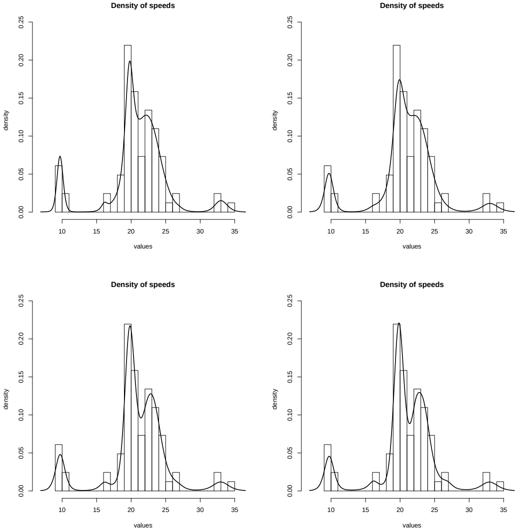
For "standard" models, you can use the `R` package `DPpackage` to perform computations.

- As an example, we analyze the `galaxy` data set: velocities (km/second) for 82 galaxies, drawn from six well-separated conic sections of the Corona Borealis region.
- The model is a location-scale DP mixture of Gaussian distributions, with a conjugate normal-inverse gamma baseline distribution:

$$
\begin{aligned}
y_i &\sim \mathrm{N}(y_i \mid \mu_i, \sigma_i^2) \\
(\mu_i, \sigma_i^2) &\sim G \\
G &\sim \mathrm{DP}(\alpha, H) \\
H &= \mathrm{N}(\mu \mid \mu_0, \sigma^2/\kappa_0)\mathrm{IGam}(\sigma^2, \nu_1, s_1)
\end{aligned}
$$

- Four different prior specifications are considered.

The Galaxy data



(1) (2) (3) (4)

The R code

```
library(DPpackage)
data(galaxy)
galaxy <- data.frame(galaxy,speeds=galaxy$speed/1000)
attach(galaxy)
state <- NULL
nburn <- 1000
nsave <- 10000
nskip <- 10
ndisplay <- 100
mcmc <- list(nburn=nburn,nsave=nsave,nskip=nskip,ndisplay=ndisplay)
# Fixing alpha, m1, and s1
prior1 <- list(alpha=1,m1=rep(0,1),psiinv1=diag(0.5,1),nu1=4,tau1=1,tau2=100)
# Fixing alpha and m1
prior2 <- list(alpha=1,m1=rep(0,1),psiinv2=solve(diag(0.5,1)),nu1=4,nu2=4, tau1=1,tau2=100)
# Fixing only alpha
prior3 <-
list(alpha=1,m2=rep(0,1),s2=diag(100000,1),psiinv2=solve(diag(0.5,1)),nu1=4,nu2=4,tau1=1,tau2=100)
#Everything is random
prior4 <- list(a0=2,b0=1,m2=rep(0,1),s2=diag(100000,1), psiinv2=solve(diag(0.5,1)),
nu1=4,nu2=4,tau1=1,tau2=100)
fit1.1 <- DPdensity(y=speeds,prior=prior1,mcmc=mcmc,state=state,status=TRUE)
fit1.2 <- DPdensity(y=speeds,prior=prior2,mcmc=mcmc,state=state,status=TRUE)
fit1.3 <- DPdensity(y=speeds,prior=prior3,mcmc=mcmc,state=state,status=TRUE)
fit1.4 <- DPdensity(y=speeds,prior=prior4,mcmc=mcmc,state=state,status=TRUE)
plot(fit1.1,ask=FALSE)
plot(fit1.2,ask=FALSE)
plot(fit1.3,ask=FALSE)
plot(fit1.4,ask=FALSE)
```

Semiparametric linear mixed effect models

- Bayesian version of Laird & Ware (1982):

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \qquad \epsilon_i \sim_{iid} \mathrm{N}(0, \sigma^2) \qquad \beta \sim \mathrm{N}(0, \Sigma)$$

with $b_i \sim \mathrm{N}(0, \Omega)$.

- A semiparametric version (Mukhopadhyay and Gelfand, 1997; Kleinman and Ibrahim, 1998)

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \qquad \epsilon_i \sim_{iid} \mathrm{N}(0, \sigma^2) \qquad \beta \sim \mathrm{N}(0, \Sigma)$$

with $b_i \sim G$ and $G \sim \mathrm{DP}\{\alpha, \mathrm{N}(0, \Omega)\}$.

- Example: The `ergoStool` dataset from the package `nlme` (Pinheiro and Bates, 2000)

$$\text{effort} \ \sim \ \text{Type (fixed)} \ + \ \text{Subject (random)}$$