Lecture 3
Hierarchical and latent class models

olivier.francois@imag.fr

April 2011

Outline

- Hierarchical models
- Mixture and latent class models
- Gibbs sampler algorithm for latent class models

Hierarchical models

- Bayesian models can be defined hierarchically.
- Hyperprior distribution $p(\psi)$
- Prior distribution $p(\theta|\psi)$
- likelihood $p(y|\theta)$
- In this case, the posterior distribution is

$$p(\theta, \psi|y) \propto p(y|\theta)p(\theta|\psi)p(\psi).$$

Mixture model

- Case 1. $y \in \mathbb{R}$
- Convex combination of densities. Let $(p_k)_{k=1,\dots,K}$, $\theta = (\theta_k)_{k=1,\dots,K}$ and $\sum_{k=1}^{K} p_k = 1$

$$p(y|\theta) = \sum_{k=1}^{K} p_k p(y|\theta_k)$$

- Gaussian mixture. For $\theta_k = (m_k, \sigma_k^2)$, we have

$$p(y|\theta_k) = N(m_k, \sigma_k^2)(y)$$

Mixture model

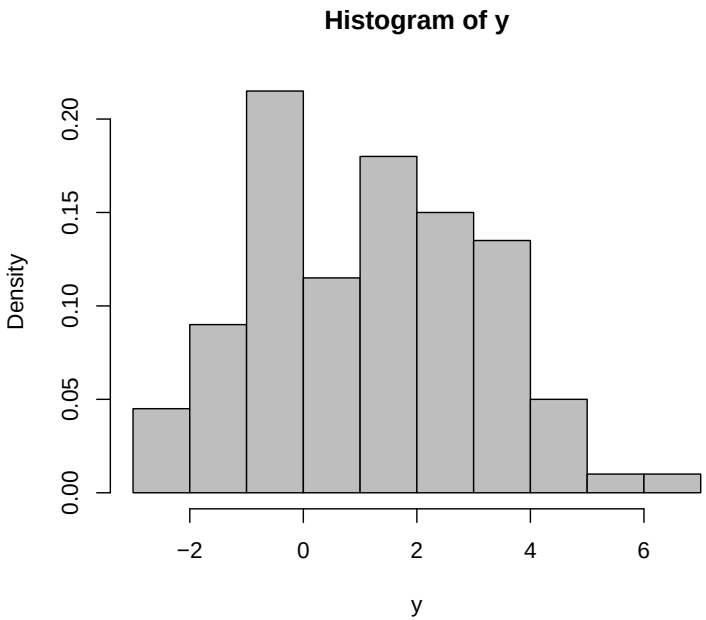
- Case 2. $y = (y_1, \ldots, y_n)$
- Sampling distribution.

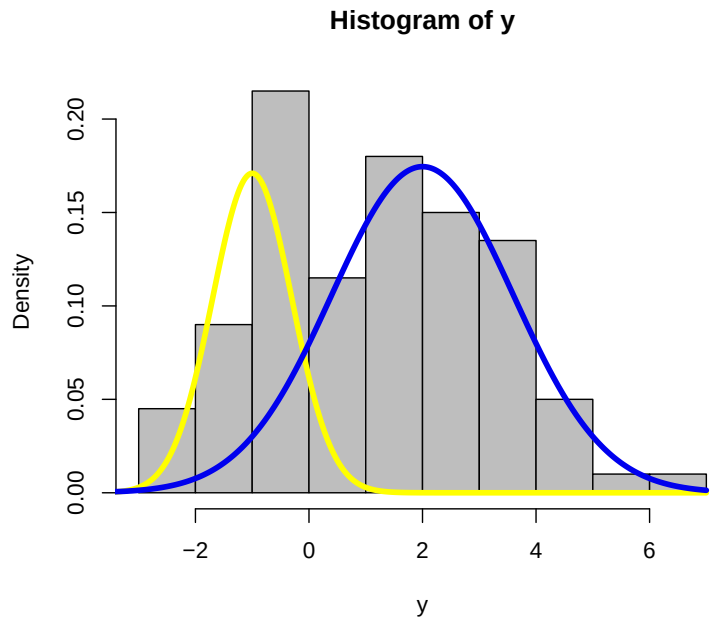
$$p(y|\theta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} p_k p(y_i|\theta_k) \right)$$

- Gaussian mixture. For $i = 1, \ldots, n$,

$$p(y_i|\theta_k) = N(m_k, \sigma_k^2)(y_i)\,.$$

Mixture model ($K = 2$)





$$p_1 = 30\% \quad p_2 = 70\%$$

Clustering problems

- How many groups in the data?
- What are the within-group means and variances?
- For a given individual, what is the assignment probability?

Mixtures as hierarchical models

- Introduce a hidden (unobserved) class label $z_i \in \{1, \ldots, K\}$ for each y_i
- For $i = 1, \ldots, n$ consider

$$p(y_i|\theta, z_i) = p(y_i|\theta_{z_i}) = N(m_{z_i}, \sigma^2_{z_i})(y_i)$$

- With this representation, we obtain

$$p(y_i|\theta) = \sum_{k=1}^{K} p(y_i|\theta_k) p(z_i = k)$$

Bayesian model for Gaussian mixtures

- Hyperprior distribution on class labels: $p(z) \propto 1$ $(p(z_i = k) = \frac{1}{K})$
- Prior distribution on parameters

$$p(\theta) = \frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2}\cdots\frac{1}{\sigma_K^2}$$

- Sampling distribution

$$p(y_i|\theta, z_i) = N(m_{z_i}, \sigma_{z_i}^2)(y_i), \quad i = 1, \ldots, n$$

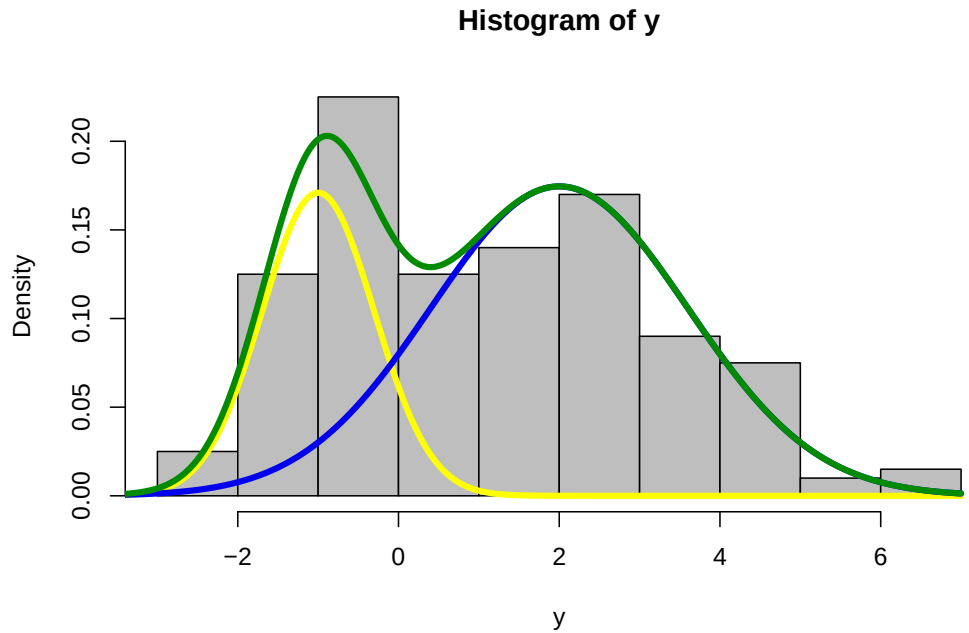
- z is a hidden variable.

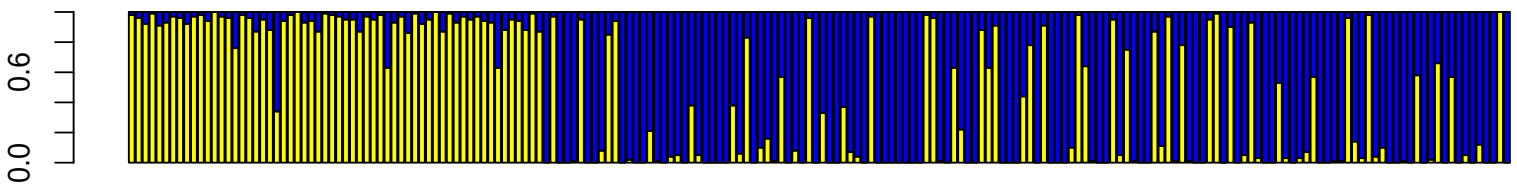
Creating an artificial data set ($n = 200$)

- $y_1, \ldots, y_{60}$ are sampled from class 1.

```
z.truth = c(rep(1,60), rep(2, 140))
m.truth = c(-1, 2)
s.truth = c(.7, 1.6)
for (i in 1:200){
y[i]=rnorm(1,m.truth[z.truth[i]],sd=s.truth[z.truth[i]])
}
```

Barplots: Marginal posterior distributions on class labels $p(z_i|y)$





Clustering as an inference problem

- Clustering consists of estimating the unobserved class label z_i for each i.
- This can achieved by computing the marginal posterior distributions, $p(z_i|y)$,
- from the joint posterior distribution

$$p(\theta, z|y) \propto p(y|\theta, z)p(\theta)p(z)$$

Gibbs sampler for the mixture model

- Consider the multidimensional parameter ($n + 2K$ dimensions)

$$(z, \theta) = (z_1, \ldots, z_n, m_1, \ldots, m_K, \sigma_1^2, \ldots, \sigma_K^2)$$

- The Gibbs sampler cycle (sweep) is

GS1. For $i = 1, \ldots, n$, update $z_i \sim p(z_i|\theta, y_i)$ (conditional independence)

GS2. For $k = 1, \ldots, K$, update $m_k \sim p(m_k|z, m_{-k}, \sigma^2)$ where

$$m_{-k} = (\ldots, m_{k-1}, m_{k+1}, \ldots)$$

GS3. For $k = 1, \ldots, K$, update $\sigma_k^2 \sim p(\sigma_k^2|z, m, \sigma_{-k}^2)$

Joint posterior distribution

- The joint posterior distribution is

$$
\begin{aligned}
p(\theta, z|y) &\propto \left(\prod_{i=1}^{n} p(y_i|\theta, z_i) p(z_i) \right) p(\theta) \\
&\propto \prod_{i=1}^{n} p(y_i|\theta, z_i) \prod_{k=1}^{K} \frac{1}{\sigma_k^2} \\
&\propto \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma_{z_i}^2}} \exp\left(-\frac{1}{2\sigma_{z_i}^2}(y_i - m_{z_i})^2\right) \prod_{k=1}^{K} \frac{1}{\sigma_k^2}
\end{aligned}
$$

Gibbs sampler GS1

- Conditional distribution on class labels (Bayes formula)

$$
\begin{aligned}
p(z_i|\theta, y_i) &= \frac{p(y_i|\theta, z_i)}{\sum_k p(y_i|\theta, k)} \\
&\propto \frac{\frac{1}{\sqrt{\sigma^2_{z_i}}} \exp(-\frac{1}{2\sigma^2_{z_i}}(y_i - m_{z_i})^2)}{\sum_{k=1}^{K} \frac{1}{\sqrt{\sigma^2_k}} \exp(-\frac{1}{2\sigma^2_k}(y_i - m_k)^2)}
\end{aligned}
$$

R script

- Use the R function `sample`

```
for (i in 1:n) {
p = exp(-(y[i] - m)^ 2/2/sigma2 )/sqrt(sigma2)
z[i] = sample(1:K, 1, prob = p ) }
```

Gibbs sampler GS2

- Conditional distribution on within-group means
- Let n_k be the current size of class k, $n_k = \#\{i : z_i = k\}$, we have (exercise)

$$p(m_k | m_{-k}, \sigma^2, z, y) = N(\bar{y}_k, \frac{\sigma_k^2}{n_k})$$

with

$$\bar{y}_k = \frac{1}{n_k} \sum_{i:z_i=k} y_i.$$

R script

- Use the R function `rnorm`

```
nk[k]=sum(z==k)
m[k]=rnorm(1,mean(y[z==k]),sd=sqrt(sigma2[k]/nk[k]))
```

Gibbs sampler GS3

- Conditional distribution on within-group variances
- Let n_k be the current size of class k, $n_k = \#\{i : z_i = k\}$, we have (exercise)

$$p(\sigma_k^2 | m, \sigma_{-k}^2, z, y) = Inv\chi^2(n_k, s_k^2)$$

with

$$s_k^2 = \frac{1}{n_k} \sum_{i:z_i=k} (y_i - m_k)^2.$$

R script

- Use the R function `rchisq`

```
sigma2[k] = sum((y[z==k]-m[k])^ 2)/rchisq(1,
nk[k])
```

Gibbs sampler for mixture model (1)

```
mcmc.mix=function(y,niter=1000,m.o=
c(0,1),sigma2.o=c(1,1)) {
n=length(y);K=length(m.o);m=m.o;s2=sigma2.o
z=NULL;p=NULL;nk = NULL
p.mcmc=NULL;m.mcmc=NULL;sigma2.mcmc=NULL;z.mcmc=NULL
for(nit in 1:niter) {
for(i in 1:n) {
p=exp(-(y[i]-m)^ 2/2/s2)/sqrt(s2)
z[i]=sample(1:K,1,prob = p) }
z.mcmc=rbind(z.mcmc,z)
```

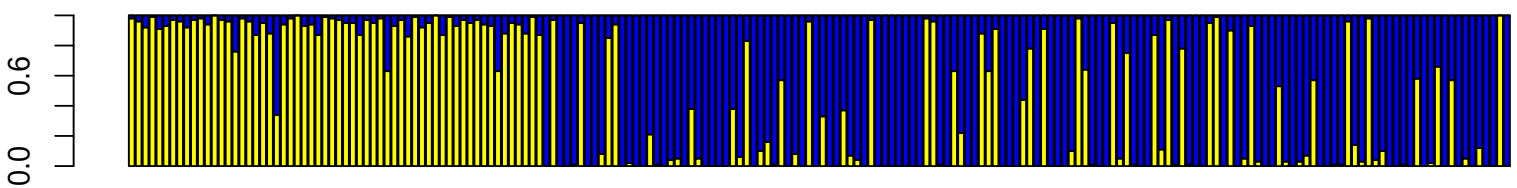
Gibbs sampler for mixture model (2)

```
for(k in 1:K) {
nk[k]=sum(z==k)
m[k]=rnorm(1,mean(y[z==k]),sd=sqrt(s2[k]/nk[k]))
s2[k]=sum((y[z==k]-m[k])^ 2)/rchisq(1,nk[k]) }
m.mcmc=rbind(m.mcmc,m)
sigma2.mcmc=rbind(sigma2.mcmc,s2)
}
return(list(z=z.mcmc,m=m.mcmc,sigma2=sigma2.mcmc))
}
```

Results for the simulated data (K = 2 classes)

- Run the program for y (30% of the total data set are from class 1)

```
obj=mcmc.mix(y,niter=300,m.o=c(2,4),sigma2.o=c(1,1))
mat = rbind(apply(obj$z[-(1:100),],MARGIN=2,
FUN=function(a)mean(a==1)),
apply(obj$z[-(1:100),],MARGIN=2,FUN=function(a)
mean(a==2)),
apply(obj$z[-(1:100),],MARGIN=2,FUN=function(a)
mean(a==3)))
barplot(mat,col=c("yellow","blue2","red"))
```



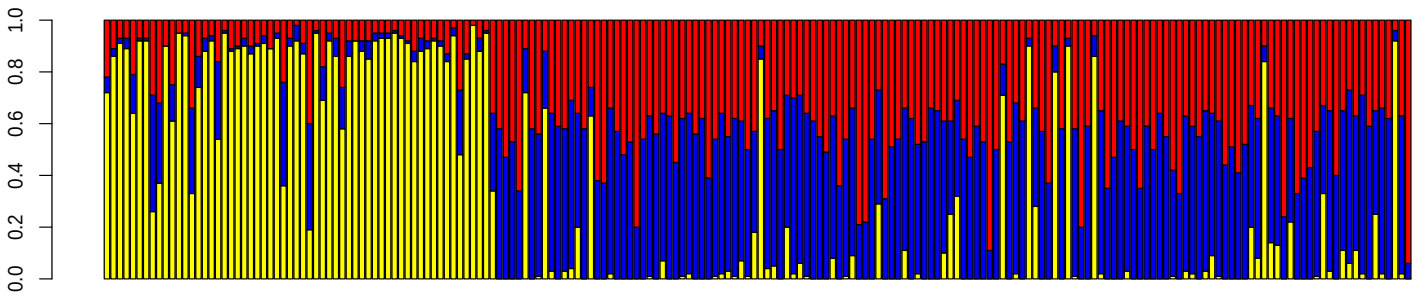
Exercises

- Display posterior distributions for the within-group means and variances.
- Estimate the number of mis-assigned data ($\approx 15\%$).

Results for the simulated data (K = 3 classes)

- Run the program for y (30% data from class 1)

```
obj=mcmc.mix.k(y,niter=200,m.o=c(-2,1,2),sigma2.o
= c(1,1,1))
mat = rbind(apply(obj$z[-(1:100),],MARGIN=2,
FUN=function(a)mean(a==1)),
apply(obj$z[-(1:100),],MARGIN=2,FUN=function(a)
mean(a==2)),
apply(obj$z[-(1:100),],MARGIN=2,FUN=function(a)
mean(a==3)))
barplot(mat,col=c("yellow","blue2","red"))
```



Exercises

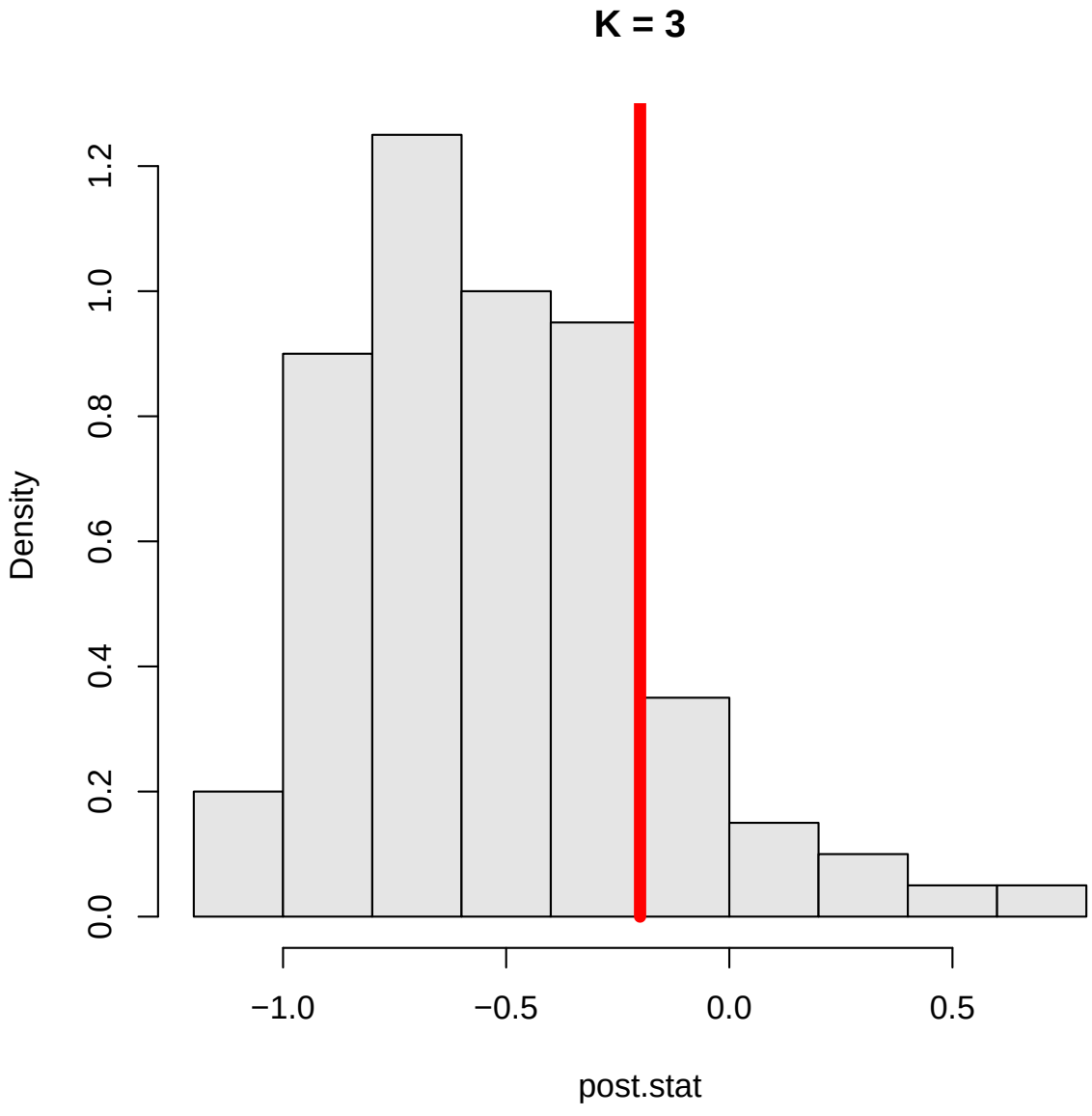
- Display posterior distributions for the within-group means and variances.
- Estimate the number of mis-assigned data ($\approx 8\%$).
- which K is correct? Show the components in the $K = 3$ model.
- Posterior predictive checks

Posterior predictive checks

- Check the 3 component model with the `kurtosis` (`skewness`) statistics

```
post.stat = NULL
for (i in 1:100){
zz = obj$z[i+100,];mm = obj$m[i+100,];ss2 =
obj$sigma2[i+100,]
yy = c(rnorm(sum(zz==1),mm[1],sqrt(ss2[1])),
rnorm(sum(zz==2),mm[2],sqrt(ss2[2])),
rnorm(sum(zz==3),mm[3],sqrt(ss2[3])))
post.stat=c(post.stat,kurtosis(yy))}
```

Posterior predictive checks



Exercises

Ex1. Implement the MCMC algorithm for mixtures of one-dimensional Gaussian distribution.

Ex2. Analyze the iris sepal length data set (`data(iris); y = iris[,1]`)

Ex3. Run the algorithm with $K = 2$ and $K = 3$.

Ex4. Which K is the best model?

Bibliography and resources

- Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis 2nd ed. Chapman & Hall, New-York.
- E. Paradis (2005) R pour les débutants. Univ. Montpellier II.
- R website: `http://cran.r-project.org/`