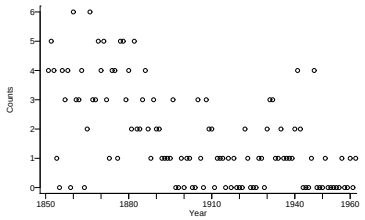
Poisson Process with Change Point

Example: British coalmining disaster data, 1851-1962

Year	Count	Year	Count	Year	Count	Year	Count
1851	4	1879	3	1907	0	1935	2
1852	5	1880	4	1908	3	1936	1
1853	4	1881	2	1909	2	1937	1
1854	1	1882	5	1910	2	1938	1
1855	0	1883	2	1911	0	1939	1
1856	4	1884	2	1912	1	1940	2
1857	3	1885	3	1913	1	1941	4
1858	4	1886	4	1914	1	1942	2
1859	0	1887	2	1915	0	1943	0
1860	6	1888	1	1916	1	1944	0
1861	3	1889	3	1917	0	1945	0
1862	3	1890	2	1918	1	1946	1
1863	4	1891	2	1919	0	1947	4
1864	0	1892	1	1920	0	1948	0
1865	2	1893	1	1921	0	1949	0
1866	6	1894	1	1922	2	1950	0
1867	3	1895	1	1923	1	1951	1
1868	3	1896	3	1924	0	1952	0
1869	5	1897	0	1925	0	1953	0
1870	4	1898	0	1926	0	1954	0
1871	5	1899	1	1927	1	1955	0
1872	3	1900	0	1928	1	1956	0
1873	1	1901	1	1929	0	1957	1
1874	4	1902	1	1930	2	1958	0
1875	4	1903	0	1931	3	1959	0
1876	1	1904	0	1932	3	1960	1
1877	5	1905	3	1933	1	1961	0
1878	5	1906	1	1934	1	1962	1



Poisson Process with Change Point

Model: Poisson process with a change point

- The distribution changes after first m observations:

$$Y_i \overset{\text{iid}}{\sim} \text{Poisson}(\theta_1) \qquad \text{for } i = 1, \ldots, m$$
$$Y_i \overset{\text{iid}}{\sim} \text{Poisson}(\theta_2) \qquad \text{for } i = m+1, \ldots, n$$

- Parameter (m, θ_1, θ_2)
- m is called a *change point*

Bayesian approach:

- Prior distributions

$$\pi(\theta_1) \sim \Gamma(a_1, b_1)$$
$$\pi(\theta_2) \sim \Gamma(a_2, b_2)$$
$$\pi(m) \sim \frac{1}{n}$$

- Conditional posterior distributions

$$\pi(\theta_1|Y, m) \sim \Gamma\Big(a_1 + \sum_{i=1}^{m} Y_i,\ m + b_1\Big)$$
$$\pi(\theta_2|Y, m) \sim \Gamma\Big(a_2 + \sum_{i=m+1}^{n} Y_i,\ n - m + b_2\Big)$$
$$\pi(m|Y, \theta_1, \theta_2) \sim c \cdot \exp\big((\theta_2 - \theta_1)\, m\big)\Big(\frac{\theta_1}{\theta_2}\Big)^{\sum_{i=1}^{m} Y_i}$$

Poisson Process with Change Point

Application of MCMC sampling

- *Step 1:* Draw

$$\theta_1^{(k)} \sim \pi\big(\theta_1 \big| Y, m^{(k-1)}\big)$$
$$\theta_2^{(k)} \sim \pi\big(\theta_2 \big| Y, m^{(k-1)}\big)$$

- *Step 2:* Draw

$$m^{(k)} \sim \pi\big(m \big| Y, \theta_1^{(k)}, \theta_2^{(k)}\big)$$

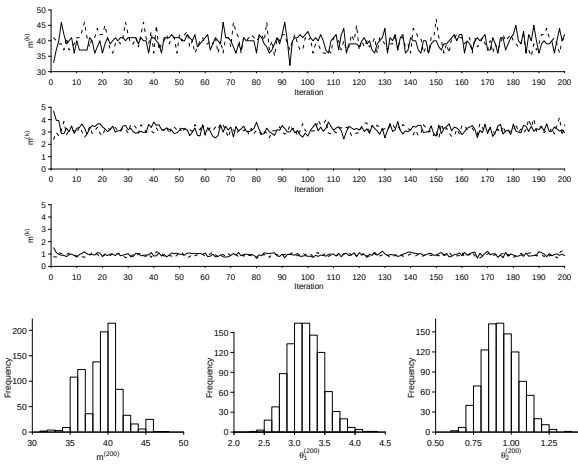
- Repeat previous two steps until stationary distributions is reached.

Implementation in *R*

```
MC<-1000                                    # number of draws (chains)
N<-200                                      # length of chains
Y<-scan("coal.txt")                         # read in data
n<-length(Y)                                # number of observations
m<-n                                        # no change point
p<-rep(0,3*MC*N)                            # array to store chains
dim(p)<-c(3,MC,N)
for (j in (1:MC)) {
  a1<-3                                     # parameter of priors
  a2<-1
  b1<-0.5
  b2<-0.5
  m<-as.integer(n*runif(1))+1
  for (i in (1:N)) {
    l1<-rgamma(1,a1+sum(Y[1:m]),m+b1)       # step 1
    l2<-rgamma(1,a2+sum(Y)-sum(Y[1:m]),n-m+b2)
    pm<-exp((l2-l1)*(1:n))*(l1/l2)^cumsum(Y)   # step 2
    pm<-pm/sum(pm)
    m<-min((1:n)[runif(1)<cumsum(pm)])
    p[1,j,i]<-m                             # save result
    p[2,j,i]<-l1
    p[3,j,i]<-l2
  }
}
```


Poisson Process with Change Point

Results:



Hierarchical Bayesian Modelling

Bayesian approach to inference

Inference is based on the posterior distribution of θ given data Y

$$\pi(\theta|Y) = \frac{f(Y|\theta)\,\pi(\theta)}{f(Y)}$$

where
- $f(Y|\theta)$ is the *likelihood function* (statistical model for data);
- $\pi(\theta)$ is the *prior distribution* of θ (quantifies uncertainty about θ);
- $f(Y) = \int f(Y|\theta)\,\pi(\theta)\,d\theta$.

Example: Binomial distribution

Suppose that X is binomially distributed with parameter θ,

$$X \sim \text{Bin}(n, \theta).$$

An appropriate prior distribution for θ is the Beta distribution

$$\theta \sim \text{Beta}(\alpha, \beta), \qquad \alpha, \beta > 0.$$

Then the posterior distribution of θ given X is again a Beta distribution with parameters $X + \alpha$ and $n - X + \beta$,

$$\theta|X \sim \text{Beta}(X + \alpha, n - X + \beta).$$

Problem: Need to specify *hyperparameters* α and β.

Idea: Specify uncertainty about hyperparameters by another level of prior distributions. For example:

$$\alpha, \beta \sim \text{Exp}(1), \quad \alpha \text{ and } \beta \text{ independent}$$

We call this kind of model a *hierarchical model*.

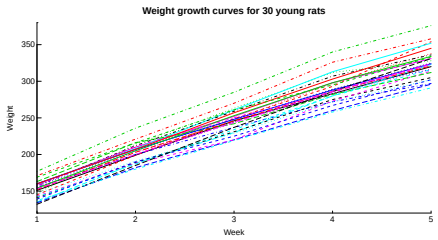
Hierarchical Bayesian Modelling

Example: Rat growth data

Data: Weight measurements of 30 young rats (weekly for five weeks)

	Week						Week				
Rat	**1**	**2**	**3**	**4**	**5**	**Rat**	**1**	**2**	**3**	**4**	**5**
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	210	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	282	312	30	153	200	244	286	324

Remarks:
- Increase in weight follows individual growth curves for each rat.
- Individual growth curves are similar in slope and variation.
- Summarize by average growth curve for population.



Hierarchical Bayesian Modelling

Data:
- Y_{ij} weight of ith rat at measurement j
- x_{ij} age (in weeks) of ith rat at measurement j
- $i = 1, \ldots, I = 30$, $j = 1, \ldots, J = 5$

Hierarchical model:

Assume individual growth curves, that is,

$$Y_{ij} \sim \mathcal{N}\big(\beta_{i0} + \beta_{i1} x_{ij}, \sigma^2\big)$$

with individual parameters $\beta_i = (\beta_{i0}, \beta_{i1})^\mathsf{T}$ distributed according to

$$\beta_i \sim \mathcal{N}\big(\beta_0, \Sigma\big).$$

Prior specifications for β_0, σ^2 and Σ:

$$\begin{aligned}
&\beta_0 \sim \mathcal{N}\big(\beta_*, \Sigma_*\big) \\
&\frac{1}{\sigma^2} \sim \Gamma\Big(\frac{\nu_*}{2}, \frac{\nu_* \tau_*^2}{2}\Big) \\
&\Sigma^{-1} \sim W\big((\rho_* R_*)^{-1}, \rho_*\big)
\end{aligned}$$

Here we take

$$\Sigma_*^{-1} = 0,\ \nu_* = 0,\ \rho_* = 2,\ R_* = \text{diag}(100, 1/10).$$

This leads to an *improper prior* (refecting vague prior information)

$$\pi(\beta_0, \sigma^2, \Sigma) \sim \frac{1}{\sigma^2} |\Sigma|^{-\frac{3+\rho_*}{2}} \exp\big(-\frac{1}{2}\text{tr}(\rho_* R_* \Sigma^{-1})\big)$$

More about the Wishart distribution: Schafer, p 150ff

Hierarchical Bayesian Modelling

Posterior distributions

The full conditional posterior distributions are:

$$\beta_i|Y, \beta_0, \Sigma, \sigma^2 \sim \mathcal{N}\big(\tfrac{1}{\sigma^2} D_i^{-1} X_i^\mathsf{T} Y_i, D_i^{-1}\big)$$

$$\beta_0|Y, \beta_1, \ldots, \beta_I, \Sigma, \sigma^2 \sim \mathcal{N}\Big(\bar{\beta}, \tfrac{1}{I}\Sigma\Big)$$

$$\frac{1}{\sigma^2}\Big|Y, \beta_1, \ldots, \beta_I, \beta_0, \Sigma \sim \Gamma\Big(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\big(Y_{ij} - \beta_{i0} - \beta_{i1} x_{ij}\big)^2\Big)$$

$$\Sigma^{-1}|Y, \beta_1, \ldots, \beta_I, \beta_0, \sigma^2 \sim W\Big(\Big[\sum_{i=1}^{I}(\beta_i - \beta_0)(\beta_i - \beta_0)^\mathsf{T} + \rho_* R_*\Big]^{-1}, I + \rho\Big)$$

where D_i is given by

$$D_i = \tfrac{1}{\sigma^2} X_i^\mathsf{T} X_i \ + \Sigma^{-1}.$$

These distributions can be used to sample from the joint posterior distribution using the *Gibbs sampler*.

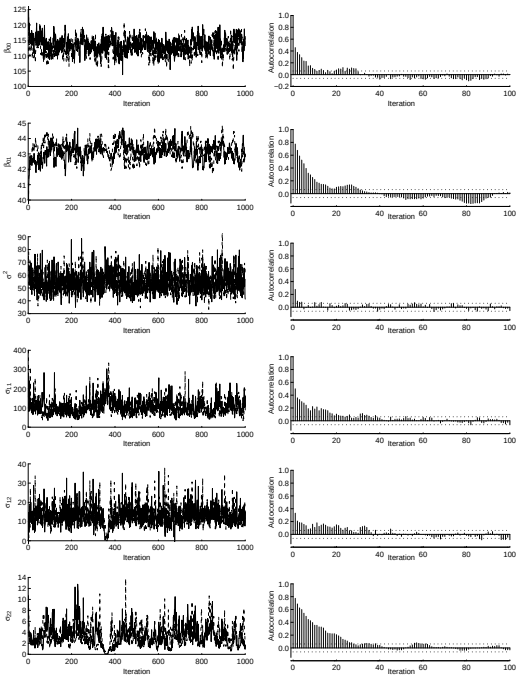
Hierarchical Bayesian Modelling

Implementation in *R*

```
# Sampling from the Wishart distribution
rwishart <- function(df, p = nrow(SqrtSigma), SqrtSigma = diag(p)) {
  if((Ident <- missing(SqrtSigma)) && missing(p))
    stop("either p or SqrtSigma must be specified")
  Z <- matrix(0, p, p)
  diag(Z) <- sqrt(rchisq(p, df:(df-p+1)))
  if(p > 1) {
    pseq <- 1:(p-1)
    Z[rep(p*pseq, pseq) + unlist(lapply(pseq, seq))] <- rnorm(p*(p-1)/2)
  }
  if(Ident)
    crossprod(Z)
  else
    crossprod(Z %*% SqrtSigma)
}
# Sampling from the multivariate normal distribution
rmultinorm<-function(n,m,S) {
  d<-ifelse(is.null(nrow(m)),length(m),nrow(m))
  m+chol(S)%*%matrix(rnorm(d*n),d,n)
}
# Rat growth data
J<-30
I<-5
Y<-matrix(scan("rats.txt"),I,J)
X<-c(rep(1,I),(1:I))
dim(X)<-c(I,2)
# Regression statistics
XTX<-t(X)%*%X
XTY<-t(X)%*%Y
# Setting of parameters
rh<-2; R<-diag(c(100,0.1))
#
MC<-2;N<-1000           #Run MC=2 chains of length N=1000
p<-rep(0,6*MC*N)        #Allocate memory for results
dim(p)<-c(6,MC,N)
#
for (j in (1:MC)) {
  S<-solve(rwishart(df=rh,SqrtSigma=chol(solve(rh*R))))
  sig<-1/rgamma(1,1000,1000)
  mu<-matrix(rnorm(2,c(200,20),c(50,5)))
  for (i in (1:N)) {
    D   <- solve(XTX/sig+solve(S))
    beta<- rmultinorm(J,D%*%((XTY)/sig+solve(S)%*%matrix(mu,2,J)),D)
    mu  <- rmultinorm(1,beta%*%matrix(1,J,1)/J,S/J)
    T   <- solve((beta-matrix(mu,2,J))%*%t(beta-matrix(mu,2,J))+rh*R)
    S   <- solve(rwishart(J+rh,SqrtSigma=chol(T)))
    sig <- 1/rgamma(1,I*J/2,1/2*sum((Y-X%*%beta)^2))
    p[,j,i]<-c(mu,sig,S[1:2,1],S[2,2])
  }
}
```

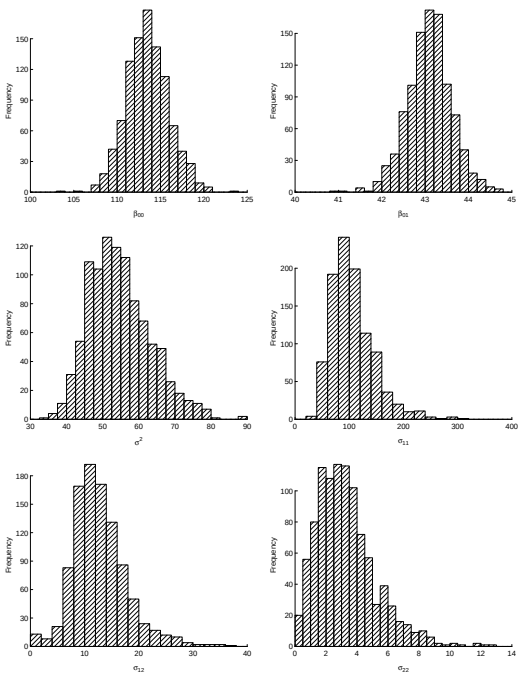

Hierarchical Bayesian Modelling

Results: Convergence and mixing of Gibbs sampler



Hierarchical Bayesian Modelling

Results: Posterior distributions



Bayesian Inference with Missing Data

Let

- $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ denote the complete data that would occur in absence of missing data where
- Y_{obs} denotes the observed values and
- Y_{mis} denotes the missing values;
- R denote the missing-data indicator.

Bayesian inference is based on the *observed-data posterior distribution*

$$\pi(\theta, \xi | Y_{\text{obs}}, R) \sim f(Y_{\text{obs}}, R|\theta, \xi)\, \pi(\theta, \xi)$$

where

$$\begin{aligned} f(Y_{\text{obs}}, R|\theta, \xi) &= \int f(Y_{\text{obs}}, y_{\text{mis}}, R|\theta, \xi)\, dy_{\text{mis}} \\ &= \int f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\, f(R|Y_{\text{obs}}, y_{\text{mis}}, \xi)\, dy_{\text{mis}} \end{aligned}$$

is the likelihood of the observed-data (i.e. Y_{obs} and R).

Assumption: Suppose that

- the missing data are missing at random, i.e. $f(R|Y) = f(R|Y_{\text{obs}})$, and
- the parameters θ and ξ are a priori independent, i.e. $\pi(\theta, \xi) = \pi(\theta)\, \pi(\xi)$.

Then inference about θ can be based on the *observed-data posterior distribution ignoring the missing-data mechanism*,

$$\pi(\theta|Y_{\text{obs}}) = \frac{f(Y_{\text{obs}}|\theta)\, \pi(\theta)}{f(Y_{\text{obs}})}$$

where

$$f(Y_{\text{obs}}) = \int f(Y_{\text{obs}}|\theta)\, \pi(\theta)\, d\theta.$$

Data Augmentation

Aim:

- Compute $\mathbb{E}\big(g(\theta)\big|Y_{\text{obs}}\big)$.
- Use MC or MCMC method for approximation

$$\mathbb{E}\big(g(\theta)\big|Y_{\text{obs}}\big) \approx \frac{1}{n}\sum_{t=1}^{n} g(\theta^{(t)}) \qquad \text{with } \theta^{(1)},\ldots,\theta^{(n)} \sim \pi(\theta|Y_{\text{obs}}).$$

Problem:

- Difficult to sample from $p(\theta|Y_{\text{obs}})$.
- Often simpler to sample from complete-data posterior $\pi(\theta|Y_{\text{obs}}, Y_{\text{mis}})$

Idea:

- Fill-in (impute) missing values to obtain complete data.
- Sample θ from the complete-data posterior distribution $\pi(\theta|Y_{\text{obs}}, Y_{\text{mis}})$.

This leads to the following iterative simulation algorithm:

Data augmentation (simplified version)

- *Imputation (I) Step:* Draw $Y_{\text{mis}}^{(t+1)}$ from $f(y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$.
- *Posterior (P) Step:* Draw $\theta^{(t+1)}$ from $\pi(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$.

Repeating the two steps from a starting value $\theta^{(0)}$ yields a Markov chain with stationary distribution $\pi(\theta, y_{\text{mis}}|Y_{\text{obs}})$.

Note: Data augmentation resembles the EM algorithm

- *E-step:* Estimate sufficient statistics (impute missing portions)
- *M-step:* Maximize complete-data likelihood (solve complete-data problem)

Data Augmentation

Example: Incomplete univariate data

Suppose that

- $Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} \text{Bin}(1,\theta)$,
- $\theta \sim \text{Beta}(a,b)$ for some fixed $a,b>0$.

Then the posterior distribution of θ is

$$\theta|Y \sim \text{Beta}\Big(a+\sum_{i=1}^{n} Y_i, b+n-\sum_{i=1}^{n} Y_i\Big).$$

Now suppose that $Y_{m+1},\ldots,Y_n$ are missing, that is, $Y_{\text{obs}} = (Y_1,\ldots,Y_m)$. It follows that

$$\theta|Y_{\text{obs}} \sim \text{Beta}\Big(a+\sum_{i=1}^{m} Y_i, b+m-\sum_{i=1}^{m} Y_i\Big).$$

Thus we can directly sample from the observed-data posterior.

Suppose we want to use data augmentation to sample from $\pi(\theta|Y_{\text{obs}})$:

- *I-step:*

$$Y_i^{(t+1)} \sim \text{Bin}\big(1,\theta^{(t)}\big),\ i=m+1,\ldots,n$$

- *P-step:*

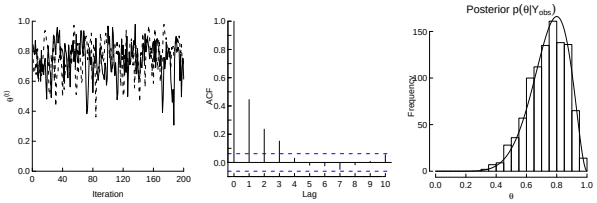
$$\theta^{(t+1)} \sim \text{Beta}\Big(a+\sum_{i=1}^{m} Y_i + \sum_{i=m+1}^{n} Y_i^{(t+1)}, b+n-\sum_{i=1}^{m} Y_i - \sum_{i=m+1}^{n} Y_i^{(t+1)}\Big)$$

Data Augmentation

Implementation in *R*

```
m<-10
n<-20
Y<-c(rbinom(m,1,0.75),rep(NA,n-m))
MC<-2;N<-1000
p<-matrix(0,MC,N)
for (j in (1:MC)) {
  th<-rbeta(1,1,1)
  for (i in (1:N)) {
    Y[(m+1):n]<-rbinom(n-m,1,th)
    th<-rbeta(1,1+sum(Y),1+n-sum(Y))
    p[j,i]<-th
  }
}
# Plotting the results
par(mfrow=c(1,3),mar=c(3,3,1,1),mgp=c(1.5,0.5,0),cex=0.8)
# (a) Time series plot of chains
plot(p[1,],type="l",xlab="Iteration",ylab=expression(theta))
lines(p[2,],lty=3)
# (b) Plot of autocorrelation function
library(ts)
acf(p[1,100:N],lag.max=50)
# (c) Histogram of posterior distribution
hist(p[1,100:N],xlab=expression(theta),main="Posterior distribution")
```

Results:



Data Augmentation

Original version of data augmentation (Tanner and Wong, 1987)

Rewrite observed-data posterior distribution as

$$\begin{aligned}\pi(\theta|Y_{\text{obs}}) &= \int \pi(\theta|Y_{\text{obs}}, y_{\text{mis}})\, f(y_{\text{mis}}|Y_{\text{obs}})\, dy_{\text{mis}} \\ &= \iint \pi(\theta|Y_{\text{obs}}, y_{\text{mis}})\, f(y_{\text{mis}}|Y_{\text{obs}}, \theta')\, \pi(\theta'|Y_{\text{obs}})\, d\theta\, dy_{\text{mis}}\end{aligned}$$

This suggests the following iterative scheme for approximating $\pi(\theta|Y_{\text{obs}})$.

Let $\pi^{(t)}(\theta|Y_{\text{obs}})$ be the current approximation of $\pi(\theta|Y_{\text{obs}})$.

- Draw $(Y_{\text{mis}}^{(1)}, \theta^{(1)}),\ldots,(Y_{\text{mis}}^{(m)}, \theta^{(m)})$ from

$$f^{(t)}(y_{\text{mis}}, \theta|Y_{\text{obs}}) = f(y_{\text{mis}}|Y_{\text{obs}}, \theta)\, p^{(t)}(\theta|Y_{\text{obs}})$$

  in two steps:

  · Draw $\theta^{(k)} \overset{\text{iid}}{\sim} \pi^{(t)}(\theta|Y_{\text{obs}})$, $k=1,\ldots,m$.

  · Draw $Y_{\text{mis}}^{(k)} \sim f(y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})$, $k=1,\ldots,m$.

  Then $Y_{\text{mis}}^{(1)},\ldots,Y_{\text{mis}}^{(m)}$ is approximately a sample from $f(y_{\text{mis}}|Y_{\text{obs}})$.

- Use Monte Carlo integration to approximate $\pi(\theta|Y_{\text{obs}})$ by

$$\pi^{(t+1)}(\theta|Y_{\text{obs}}) = \frac{1}{m}\sum_{k=1}^{m} \pi(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(k)}).$$

For $m=1$, this data augmentation algorithm reduces to the Gibbs sampler on the previous slide.

Data Augmentation

Example: Cholesterol levels of heart-attack patients

Data:

- Serum-cholesterol levels for $n = 28$ patients treated for heart attacks.
- Cholesterol levels were measured for all patients 2 and 4 days after the attack.
- For 19 of the 28 patients, an additional measurement was taken 14 days after the attack.
- See also Schafer, sections 5.3.6 and 5.4.3.

Id	Y_1	Y_2	Y_3	Id	Y_1	Y_2	Y_3
1	270	218	156	15	294	240	264
2	236	234	—	16	282	294	—
3	210	214	242	17	234	220	264
4	142	116	—	18	224	200	—
5	280	200	—	19	276	220	188
6	272	276	256	20	282	186	182
7	160	146	142	21	360	352	294
8	220	182	216	22	310	202	214
9	226	238	248	23	280	218	—
10	242	288	—	24	278	248	198
11	186	190	168	25	288	278	—
12	266	236	236	26	288	248	256
13	206	244	—	27	244	270	280
14	318	258	200	28	236	242	204

Data Augmentation

Bayesian model

Data model: $Y = (Y_1, \ldots, Y_n)^\mathsf{T}$ ($n \times p$ matrix) with

$$Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma)$$

Prior distribution (improper prior):

$$\pi(\mu, \Sigma) \sim |\Sigma|^{-\frac{p+1}{2}} = |\Sigma|^{-2}$$

Full conditionals of *posterior distribution*

$$\mu | Y, \Sigma \sim \mathcal{N}\big(\bar{Y}, \Sigma/n\big)$$
$$\Sigma^{-1} | Y, \mu \sim W\Big(\big[(Y-\mu)^\mathsf{T}(Y-\mu)\big]^{-1}, n\Big)$$

Data augmentation algorithm

- *I-step*

$$Y_{i3} \sim \mathcal{N}\big(\mu^{(t)}_{3|12}, \sigma^{(t)}_{33|12}\big)$$

where

$$\mu^{(t)}_{3|12} = \mu^{(t)}_3 + \begin{pmatrix} \sigma^{(t)}_{31} & \sigma^{(t)}_{31} \end{pmatrix} \begin{pmatrix} \sigma^{(t)}_{11} & \sigma^{(t)}_{12} \\ \sigma^{(t)}_{21} & \sigma^{(t)}_{22} \end{pmatrix}^{-1} \begin{pmatrix} Y_{i1} - \mu^{(t)}_1 \\ Y_{i2} - \mu^{(t)}_2 \end{pmatrix}$$

$$\sigma^{(t)}_{33|12} = \sigma^{(t)}_{33} - \begin{pmatrix} \sigma^{(t)}_{31} & \sigma^{(t)}_{32} \end{pmatrix} \begin{pmatrix} \sigma^{(t)}_{11} & \sigma^{(t)}_{12} \\ \sigma^{(t)}_{21} & \sigma^{(t)}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \sigma^{(t)}_{13} \\ \sigma^{(t)}_{23} \end{pmatrix}$$

- *P-step*

$$\mu^{(t+1)} \sim \mathcal{N}\big(\bar{Y}, \Sigma^{(t)}/n\big)$$
$$\Sigma^{(t+1)} \sim W^{-1}\Big(\big[(Y-\mu^{(t+1)})^\mathsf{T}(Y-\mu^{(t+1)})\big]^{-1}, n\Big)$$

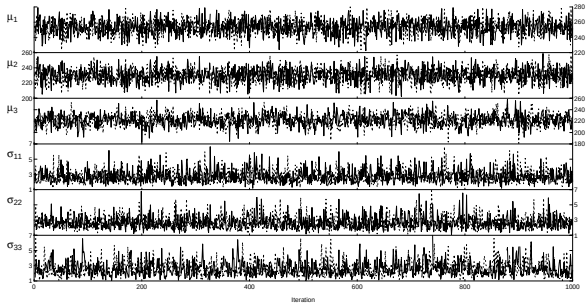
Data Augmentation

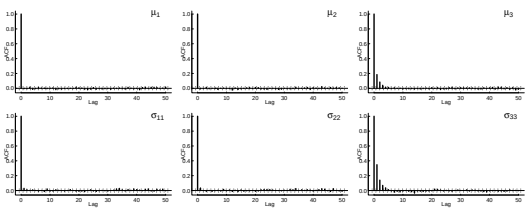
Implementation in *R*

```
Y<-t(matrix(scan("heart.txt"),3,28))
R<-ifelse(is.na(Y),0,1)
#
MC<-5
N<-1000
p<-rep(0,9*N*MC)
dim(p)<-c(9,MC,N)
#
for (j in (1:MC)) {
  S<-solve(rwishart((nrow(Y)-1)/2,3))
  m<-rnorm(3,mean(Y[R==1]),sd(Y[R==1]))
  for (i in (1:N)) {
    #Imputation step
    Y[R[,3]==0,3]<-rnorm(sum((1-R[,3])[R[,3]==0]),
      m[3]+S[3,-3]%*%solve(S[-3,-3])%*%(t(Y[R[,3]==0,1:2])-m[1:2]),
      sqrt(S[3,3]-S[3,-3]%*%solve(S[-3,-3])%*%S[-3,3]))
    #Posterior step
    m<-t(Y)%*%rep(1,nrow(Y))/nrow(Y)+t(chol(S))%*%rnorm(3)/sqrt(nrow(Y))
    S<-solve(rwishart((nrow(Y)-1),
      SqrtSigma=chol(solve((t(Y)-m[1:3])%*%t(t(Y)-m[1:3])))))
    p[,j,i]<-c(m,S[1:3,1],S[2:3,2],S[3,3])
  }
}
```


Data Augmentation

Results: Convergence of chains





- Fast convergence to stationary distribution
- Autocorrelation decreases rapidly (values 10 steps apart are approximately independent)
- Chains exhibit good mixing

Data Augmentation

Results: Variables of interest

- Average cholesterol level at 14 days

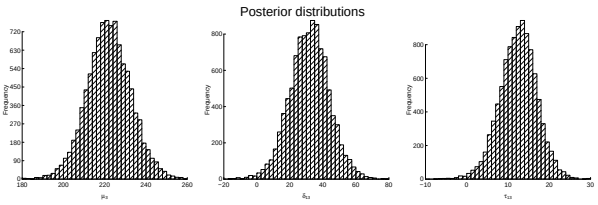
  μ_3

- Average decrease in cholesterol level from day 2 to day 14

  $\delta_{13} = \mu_1 - \mu_3$

- Relative percentage decrease in average cholesterol level from day 2 to day 14

  $\tau_{13} = \dfrac{100 \cdot (\mu_1 - \mu_3)}{\mu_1}$



Posterior means and 95% posterior intervals:

μ_3	δ_{13}	τ_{13}
222.07	31.84	12.46
[200.79, 243.68]	[8.02, 55.55]	[3.26, 21.05]

Allele Frequency Estimation

Example: ABO blood types

- ABO genetic locus exhibits three alleles: A, B, and O
- Four phenotypes: A, B, AB, and O

Genotype	A/A	A/O	A/B	B/B	B/O	O/O
Phenotype	A	A	AB	B	B	O

- *Data:* Observed counts of four phenotypes A, B, AB, and O

$\boldsymbol{n_A}$	$\boldsymbol{n_B}$	$\boldsymbol{n_{AB}}$	$\boldsymbol{n_O}$	$\boldsymbol{n}$
186	38	13	284	521

- *Aim:* Estimate frequencies p_A, p_B, and p_O of alleles A, B, and O

Modelling:

- Observed data: N_A, N_B, N_{AB}, N_O
- Complete data: N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}, N_O
- According to the Hardy-Weinberg law, the genotype frequencies are

Genotype	A/A	A/O	A/B	B/B	B/O	O/O
Frequency	p_A^2	$2p_Ap_O$	$2p_Ap_B$	p_B^2	$2p_Bp_O$	p_O^2

- Genotype counts $N = (N_{AA}, N_{AO}, N_{AB}, N_{BB}, N_{BO}, N_O)$ are jointly multinomially distributed.

Allele Frequency Estimation

- The complete data $N = (N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}, N_O)$ are multinomially distributed,

  $$N \sim M\big(n, p_A^2, 2p_Ap_O, p_B^2, 2p_Bp_O, 2p_Ap_B, p_O^2\big).$$

- The conjugate prior is the Dirichlet distribution,

  $$(p_A, p_B, p_O) \sim D\big(\alpha_A, \alpha_B, \alpha_O\big), \qquad \alpha_A, \alpha_B, \alpha_O > 0.$$

  The density of the Dirichlet distribution is given by

  $$p(p_A, p_B, p_O) = \frac{\Gamma(\alpha_A + \alpha_B + \alpha_C)}{\Gamma(\alpha_A)\Gamma(\alpha_B)\Gamma(\alpha_O)} p_A^{\alpha_A - 1} p_B^{\alpha_B - 1} p_O^{\alpha_O - 1}.$$

- The posterior distribution is again a Dirichlet distribution

  $$p_A, p_B, p_O | N \sim D\big(\alpha'_A, \alpha'_B, \alpha'_O\big)$$

  Data augmentation P-step

  with parameters

  $$\alpha'_A = \alpha_A + 2N_{AA} + N_{AO} + N_{AB}$$
  $$\alpha'_B = \alpha_B + 2N_{BB} + N_{BO} + N_{AB}$$
  $$\alpha'_O = \alpha_O + 2N_O + N_{AO} + N_{BO}$$

- Given the observed data $N_{\text{obs}} = (N_A, N_B, N_{AB}, N_O)$, the missing data $N_{\text{mis}} = (N_{AA}, N_{AO}, N_{BB}, N_{BO})$ are binomially distributed,

  $$N_{AA} \sim \text{Bin}\Big(N_A, \frac{p_A^2}{p_A^2 + 2p_Ap_O}\Big)$$
  $$N_{AO} = N_A - N_{AA}$$
  $$N_{BB} \sim \text{Bin}\Big(N_B, \frac{p_B^2}{p_B^2 + 2p_Bp_O}\Big)$$
  $$N_{BO} = N_B - N_{BB}$$

  Data augmentation I-step

Allele Frequency Estimation

Implementation in *R*

```
N<-c(186,38,13,284)                    # Data
a<-1;b<-1;c<-1                         # Prior parameters (uniform prior)
#
MC<-5
T<-1000
p<-rep(0,3*T*MC)                       # Array for parameters
dim(p)<-c(3,MC,T)
d<-rep(0,4*T*MC)                       # Array for imputed values
dim(d)<-c(4,MC,T)
for (j in (1:MC)) {                    # Loop over chains
  pa<-rbeta(1,a,b+c)                   # Starting values from
  pb<-(1-pa)*rbeta(1,b,c)              # prior distributions
  po=1-pa-pb
  for (i in (1:T)) {                   # Gibbs iterations
    # Imputation step
    Naa<-rbinom(1,N[1],pa^2/(pa^2+2*pa*po))
    Nao<-N[1]-Naa
    Nbb<-rbinom(1,N[2],pb^2/(pb^2+2*pb*po))
    Nbo<-N[2]-Nbb
    # Posterior step
    pa<-rbeta(1,a+2*Naa+Nao+N[3],b+c+2*N[2]+2*N[4]+N[3]+Nao)
    pb<-(1-pa)*rbeta(1,b+2*Nbb+Nbo+N[3],c+2*N[4]+Nao+Nbo)
    po<-1-pa-pb
    p[,j,i]<-c(pa,pb,po)
    d[,j,i]<-c(Naa,Nao,Nbb,Nbo)
  }
}
```

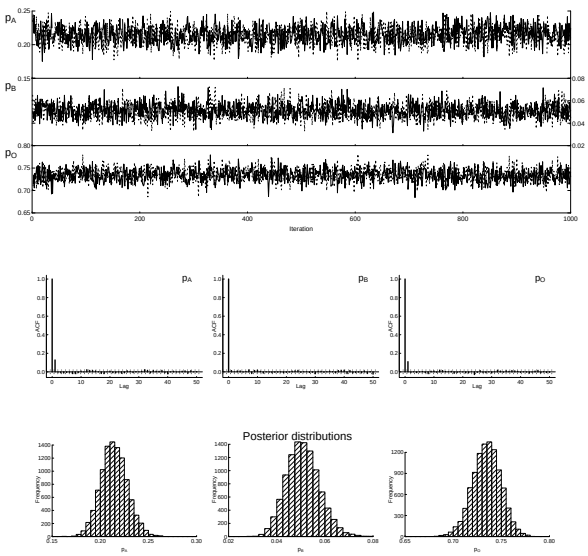
Sampling from the Dirichlet distribution

Suppose that $(p_1, \ldots, p_n) \sim D(\alpha_1, \ldots, \alpha_n)$. Then

$$\begin{aligned}
p_1 &\sim \text{Beta}(\alpha_1, \alpha_2 + \ldots + \alpha_n) \\
p_2|p_1 &\sim (1 - p_1)\text{Beta}(\alpha_2, \alpha_3 + \ldots + \alpha_n) \\
p_3|p_1, p_2 &\sim (1 - p_2 - p_3)\text{Beta}(\alpha_3, \alpha_4 + \ldots + \alpha_n) \\
&\vdots \\
p_{n-1}|p_1, \ldots, p_{n-2} &\sim (1 - p_1 - \ldots - p_{n-2})\text{Beta}(\alpha_{n-1}, \alpha_n) \\
p_n &= (1 - p_1 - \ldots - p_{n-1})
\end{aligned}$$

Allele Frequency Estimation

Results: Convergence of chains



Posterior means and 95% posterior intervals:

p_A	p_B	p_O
0.21	0.05	0.74
[0.19, 0.24]	[0.038, 0.065]	[0.71, 0.77]

Allele Frequency Estimation

Rao-Blackwell Theorem Suppose $S(\theta)$ is an unbiased estimator for some scalar quantity $s(\theta)$ and T is a sufficient statistic. Then $S^* = \mathbb{E}(S|T)$ is also unbiased and has smaller variance than S,

$$\mathrm{var}\big(\mathbb{E}(S|T)\big) \le \mathrm{var}\big(S\big).$$

Example: The direct MCMC estimator for the allele frequency p_A,

$$\hat{p}_A = \frac{1}{T}\sum_{t=1}^{T} p_A^{(t)}$$

is unbiased for $\mathbb{E}(p_A|N_{\mathrm{obs}})$. Since $N = (N_{\mathrm{obs}}, N_{\mathrm{mis}})$ is a sufficient statistic, the Rao-Blackwell Theorem suggests to use the alternative estimator

$$\hat{p}_A^* = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big(p_A\big|N_{\mathrm{obs}}, N_{\mathrm{mis}}^{(t)}\big).$$

From the conditional distribution of p_A given the complete data N, we obtain

$$\mathbb{E}\big(p_A\big|N_{\mathrm{obs}}, N_{\mathrm{mis}}\big) = \frac{1 + 2\,N_{AA} + N_{AO} + N_{AB}}{3 + 2\,n}$$

This leads to the following *Rao-Blackwellized estimates* for the allele frequencies:

$$\hat{p}_A^* = \frac{1}{T}\sum_{t=1}^{T} \frac{\alpha_A + 2\,N_{AA}^{(T)} + N_{AO}^{(T)} + N_{AB}}{\alpha_A + \alpha_B + \alpha_O + 2\,n}$$

$$\hat{p}_B^* = \frac{1}{T}\sum_{t=1}^{T} \frac{\alpha_B + 2\,N_{BB}^{(T)} + N_{BO}^{(T)} + N_{AB}}{\alpha_A + \alpha_B + \alpha_O + 2\,n}$$

$$\hat{p}_O^* = \frac{1}{T}\sum_{t=1}^{T} \frac{\alpha_O + 2\,N_O + N_{AO}^{(T)} + N_{BO}^{(T)}}{\alpha_A + \alpha_B + \alpha_O + 2\,n}$$

Estimates of posterior means (with standard deviations)

Parameter	*Direct estimate*		*Rao-Blackwellized estimate*	
p_A	0.214	(0.00043)	0.214	(0.00015)
p_B	0.051	(0.00023)	0.051	(0.00003)
p_O	0.735	(0.00046)	0.735	(0.00015)

Allele Frequency Estimation

Similarly, we can estimate the posterior cumulative distribution function by

$$\hat{\mathbb{P}}(p_A|N_{\mathrm{obs}}) = \frac{1}{t=T}\sum_{1}^{T} \mathbb{P}\big(p_A \le p|N_{\mathrm{obs}}, N_{\mathrm{mis}}^{(t)}\big),$$

and the posterior density (which is not that easy to see) by

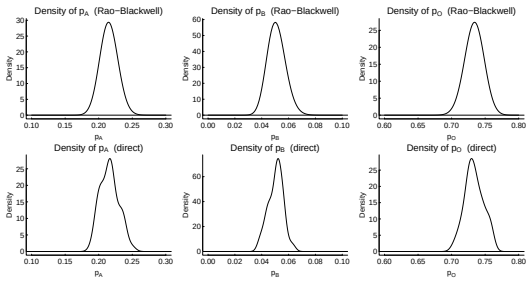
$$\hat{p}(p_A|N_{\mathrm{obs}}) = \frac{1}{T}\sum_{t=1}^{T} \pi\big(p_A|N_{\mathrm{obs}}, N_{\mathrm{mis}}^{(t)}\big),$$

where the sums are computed using

$$p_A|N_{\mathrm{obs}}, N_{\mathrm{mis}}^{(t)} \sim \mathrm{Beta}(\alpha_A^{(t)}, \alpha_B^{(t)} + \alpha_O^{(t)})$$

where $\alpha_A^{(t)}$, $\alpha_B^{(t)}$, and $\alpha_O^{(t)}$ are of the same form as α_A', α_B', and α_O', respectively.

Direct and Rao-Blackwellized density estimates for the posterior distributions:



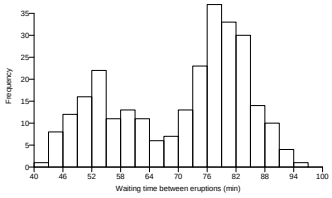
Posterior intervals: From the estimate for the posterior cumulative distribution function, we can derive a 95% posterior interval for the parameter p_A.
The following table gives results for $T = 50$ (after a burn-in period):

	Direct estimate			*Rao-Blackwellized estimate*		
Parameter	Mean	SD	95% interval	Mean	SD	95% interval
p_A	0.214	0.00191	[0.186, 0.238]	0.213	0.00060	[0.188, 0.240]
p_B	0.051	0.00083	[0.041, 0.062]	0.051	0.00015	[0.038, 0.065]
p_O	0.735	0.00202	[0.712, 0.768]	0.735	0.00060	[0.707, 0.763]

Gaussian Mixtures

Example: Old Faithful

Data: 272 waiting times between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA



Model: Mixture of two Gaussian populations (short/long waiting times):

$$f_Y(y|\theta) = \pi\,\frac{1}{\sigma_1}\,\varphi\Big(\frac{x-\mu_1}{\sigma_1}\Big) + (1-\pi)\,\frac{1}{\sigma_2}\,\varphi\Big(\frac{x-\mu_2}{\sigma_2}\Big)$$

with parameter $\theta = (\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^{\mathsf{T}}$.

- With probability π an observationY_i is drawn from a normal population with mean μ_1 and standard deviation σ_1.
- With probability $1-\pi$ an observationY_i is drawn from a normal population with mean μ_2 and standard deviation σ_2.

Idea: If we knew the group which each observation belongs to, we could simply fit a normal distribution to each group.

Missing data: Group indicator

$$Z_i = \begin{cases} 1 & Y_i \text{ belongs to group of long waiting times} \\ 0 & Y_i \text{ belongs to group of short waiting times} \end{cases}$$

Z_i is Bernoulli distributed with parameter π: $Z_i \overset{\text{iid}}{\sim} \mathrm{Bin}(1, \pi)$

Gaussian Mixtures

- The complete data (Y, Z) are distributed according to

$$f(Y, Z|\theta) = \prod_{i=1}^{n} \left[\frac{\pi}{\sigma_1}\varphi\Big(\frac{Y_i-\mu_1}{\sigma_1}\Big)\right]^{Z_i}\left[\frac{1-\pi}{\sigma_2}\varphi\Big(\frac{Y_i-\mu_2}{\sigma_2}\Big)\right]^{1-Z_i}$$

  where $\theta = (\pi, \mu_1, \mu_2, \sigma_1, \sigma_2)$.

- We adapt an improper noninformative prior

$$\pi(\theta) \sim \big[\pi(1-\pi)\big]^{-\frac{1}{2}}\sigma^{-2}.$$

  Jeffrey's prior: If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y|\theta)$, then a noninformative prior is given by $\pi(\theta) \sim |I(\theta)|^{\frac{1}{2}}$.

- The *full conditional posterior* distributions are

$$\pi \sim \text{Beta}\Big(\frac{1}{2}+N_1, \frac{1}{2}+N_2\Big) \qquad N_1 = \sum_{i=1}^{n} Z_i \text{ and } N_2 = n - N_1$$

$$\mu_1 \sim \mathcal{N}\Big(\frac{1}{N_1}\sum_{i=1}^{n} Y_i\, Z_i, \frac{\sigma_1^2}{N_1}\Big)$$

$$\mu_2 \sim \mathcal{N}\Big(\frac{1}{N_2}\sum_{i=1}^{n} Y_i\,(1-Z_i), \frac{\sigma_2^2}{N_2}\Big)$$

$$\sigma_1^{-2} \sim \Gamma\Big(\frac{1}{2}\big(N_1-1\big), \frac{1}{2}\sum_{i=1}^{n} Z_i\,(Y_i-\mu_1)^2\Big)$$

$$\sigma_2^{-2} \sim \Gamma\Big(\frac{1}{2}\big(N_2-1\big), \frac{1}{2}\sum_{i=1}^{n} (1-Z_i)\,(Y_i-\mu_1)^2\Big)$$

Data augmentation P-step

- Given the observed data Y, the missing data Z are binomially distributed,

$$Z_i|Y_i, \theta \sim \text{Bin}(1, \pi_i)$$

Data augmentation I-step

  where

$$\pi_i = \frac{\pi\,\frac{1}{\sigma_1}\varphi\Big(\frac{x-\mu_1}{\sigma_1}\Big)}{\pi\,\frac{1}{\sigma_1}\varphi\Big(\frac{Y_i-\mu_1}{\sigma_1}\Big) + (1-\pi)\,\frac{1}{\sigma_2}\varphi\Big(\frac{Y_i-\mu_2}{\sigma_2}\Big)}.$$

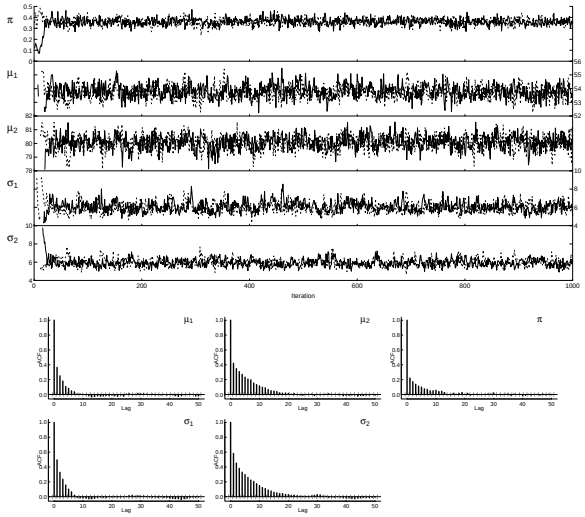
Gaussian Mixtures

Implementation in *R*

```
data(faithful)
Y<-faithful$waiting
n<-length(Y)
#
MC<-2
N<-10000
p<-rep(0,5*N*MC)
dim(p)<-c(5,MC,N)
for (j in (1:MC)) {
  pmix<-rbeta(1,0.5,0.5)
  m<-c(mean(Y),mean(Y))
  s<-c(sd(Y),sd(Y))
  for (i in (1:N)) {
    Z<-rbinom(n,1,pmix*dnorm(Y,m[1],s[1])/
       (pmix*dnorm(Y,m[1],s[1])+(1-pmix)*dnorm(Y,m[2],s[2])))
    pmix<-rbeta(1,sum(Z)+1/2,n-sum(Z)+1/2)
    if (pmix>1/2) {    # Restrict to pmix<1/2
      pmix<-1-pmix
      Z<-1-Z
    }
    m[1]<-rnorm(1,mean(Y[Z==1]),s[1]/sqrt(sum(Z)))
    m[2]<-rnorm(1,mean(Y[Z==0]),s[2]/sqrt(sum(1-Z)))
    s[1]<-1/sqrt(rgamma(1,(sum(Z)-1)/2,sum((Y[Z==1]-m[1])^2)/2))
    s[2]<-1/sqrt(rgamma(1,(sum(1-Z)-1)/2,sum((Y[Z==0]-m[2])^2)/2))
    p[,j,i]<-c(pmix,m,s)
  }
}
```


Gaussian Mixtures

Results: Convergence of chains



- Fast convergence
- Good mixing
- Moderate autocorrelation (independence for lags ≥ 30)

Estimation results:

π	0.3609	[0.2996, 0.4239]
μ_1	54.63	[53.23, 56.10]
μ_2	80.07	[79.01, 81.05]
σ_1	6.00	[5.02, 7.24]
σ_2	5.96	[5.23, 6.86]