MAS8303: Modern Bayesian Inference

Malcolm Farrow

November 2010

Lecture 4.1: Mixtures

Two reasons why we might want to do this:

1. We might believe that there really are two or more sub-populations.
2. Using a mixture distribution allows more flexibility in the sampling model.

Lecture 4.1: Mixtures

Consider a simple two-component mixture model. Our sampling model for observation Y_i has pdf

$$f(y_i;\ \pi, \theta_1, \theta_2) = \pi f_1(y_i; \theta_1) + (1 - \pi) f_2(y_i; \theta_2).$$

Here $f_j(y; \theta_j)$ is the pdf for component j and depends on parameters θ_j. The component membership probabilities are π and $1 - \pi$, with $0 \leq \pi \leq 1$.
Suppose that we have n independent (given the parameters) observations $y_1, \ldots, y_n$. The likelihood is

$$L = \prod_{i=1}^{n} \left\{ \pi f_1(y_i; \theta_1) + (1 - \pi) f_2(y_i; \theta_2) \right\}. \qquad (4.1)$$

This has a rather complicated form. For example, it is a polynomial of degree n in π.

Lecture 4.1: Mixtures

More generally we could have J components with

$$f(y_i;\ \underline{\pi}, \Theta) = \sum_{j=1}^{J} \pi_j f_j(y_i; \theta_j), \qquad (4.2)$$

where $\sum_{j=1}^{J} \pi_j = 1$ and $\pi_j \geq 0$ for $j = 1, \ldots, J$. In this case the likelihood is

$$L = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} \pi_j f_j(y_i; \theta_j) \right\}. \qquad (4.3)$$

This could be very complicated.

Lecture 4.1: Mixtures

We can make things much simpler by introducing a group-membership variable which is unobserved. The values form auxiliary data so this is an example of data augmentation.

Lecture 4.1: Mixtures

We introduce, for observation i, an auxiliary variable c_i, which can take the values $1, \ldots, J$. Then, given that $c_i = j$, the conditional pdf for observation i is simply $f_j(y_i;\ \theta_j)$. The corresponding conditional likelihood is then just

$$L_c = \prod_{i=1}^{n} \pi_{c_i} f_{c_i}(y_i;\ \theta_{c_i}).$$

Lecture 4.1: Mixtures

Now we give c_i a multinomial (or "categorical") distribution, in which $\Pr(c_i = j) = \pi_j$. We give the parameters $\underline{\pi} = (\pi_1, \ldots, \pi_J)^T$ and $\Theta = \{\theta_1, \ldots, \theta_J\}$ a suitable prior distribution. Then, by "integrating out", i.e. "averaging over", $c_1, \ldots, c_n$, we obtain the correct posterior distribution.

Lecture 4.1: Mixtures

The joint probability (density) that $c_i = j$ and $Y_i = y_i$ is

$$f(y_i, c_i = j;\ \underline{\pi}, \Theta) = \pi_j f_j(y_i;\ \theta_j).$$

To find the marginal probability density of y_i we sum over j and obtain (4.2) as required.

Lecture 4.1: MCMC and label-switching : MCMC

Once we have the model set up with the auxiliary variables $c_1, \ldots, c_n$ as above, we have a prior distribution with density $f_0(\Theta, \underline{\pi})$ for the parameters and we have initial values for the unknowns, Θ, $\underline{\pi}$, $c_1, \ldots, c_n$, then we can proceed with MCMC as follows.

Lecture 4.1: MCMC and label-switching : MCMC

1. Sample a new value for Θ. The fcd density is proportional to

$$f_0(\Theta, \underline{\pi}) \prod_{j=1}^{J} L_{c,j}$$

where

$$L_{c,j} = \prod_{i \in C_j} f_j(y_i;\ \theta_j)$$

and $i \in C_j$ if $c_i = j$. That is C_j is the set of observations currently assigned to component j. We might well have $f_0(\Theta, \underline{\pi}) = f_{0,\theta}(\Theta) f_{0,\pi}(\underline{\pi})$ in which case the fcd density is proportional to

$$f_0(\Theta) \prod_{j=1}^{J} L_{c,j}$$

Lecture 4.1: MCMC and label-switching : MCMC

2. Sample a new value for $\underline{\pi}$. The fcd density is proportional to

$$f_0(\Theta, \underline{\pi}) \prod_{j=1}^{J} \pi_j^{n_j}$$

where n_j is the number of observations currently assigned to component j. If $f_0(\Theta, \underline{\pi}) = f_{0,\theta}(\Theta) f_{0,\pi}(\underline{\pi})$ then the fcd density is proportional to

$$f_{0,\pi}(\underline{\pi}) \prod_{j=1}^{J} \pi_j^{n_j}.$$

A popular choice for $f_{0,\pi}(\underline{\pi})$ would be a Dirichlet density. In this case the fcd is also a Dirichlet distribution. Sampling from a Dirichlet distribution is quite easy.

Lecture 4.1: MCMC and label-switching : MCMC

3. Sample a new value for each of $c_1, \ldots, c_n$. The fcd is a categorical distribution with

$$\Pr(c_i = j) \propto \pi_j f_j(y_i;\ \theta_j).$$

4. Repeat.

Lecture 4.1: MCMC and label-switching : Label-switching

Consider the likelihood (4.1).
Suppose that both component distributions are of the same family so that the likelihood is

$$L = \prod_{i=1}^{n} \{\pi f_y(y_i; \theta_1) + (1 - \pi) f_y(y_i; \theta_2).\}$$

Suppose that we "switch the labels" and write

$$\tilde{L} = \prod_{i=1}^{n} \left\{\tilde{\pi} f_y(y_i; \tilde{\theta}_1) + (1 - \tilde{\pi}) f_y(y_i; \tilde{\theta}_2)\right\}$$

where $\tilde{\pi} = 1 - \pi$, $\tilde{\theta}_1 = \theta_2$ and $\tilde{\theta}_2 = \theta_1$.

Clearly $L = \tilde{L}$. The likelihood is therefore bimodal and, in fact, the modes match each other. If the prior does not strongly favour one mode over the other then the posterior distribution will also be bimodal.

Lecture 4.1: MCMC and label-switching : Label-switching

- Constraints – eg order constraints – on parameters.
- Mixtures with unknown numbers of components.

Lecture 4.1: Multivariate mixtures

It is, of course, possible to make a mixture model where the observation $\underline{y}$ is multivariate. For example, we might make several measurements on each of a sample of birds belonging to one species with the idea that there might be two or more subspecies. In two dimensions we might expect a plot of observations y_1 against y_2 to reveal "clusters" of observations.

Lecture 4.1: Continuous mixtures

As well as the finite mixtures described above it is possible to have a mixture model with an infinite number of components. It is also possible to have a *continuous mixture*. In a continuous mixture model, instead of (4.2), we have, for example,

$$f(y_i) = \int_{\Omega} f_\theta(\theta) f_y(y_i; \theta, \lambda_i). \qquad (4.4)$$

Lecture 4.1: Continuous mixtures

$$f(y_i) = \int_{\Omega} f_\theta(\theta) f_y(y_i; \theta, \lambda_i). \qquad (4.4)$$

Here θ is a parameter with a continuous distribution specified by the *mixing density* $f_\theta(\theta)$. The range of values of θ is denoted by Ω. There may be other parameters which do not vary in this way and these are denoted by λ_i.

Lecture 4.1: Continuous mixtures

We saw an example of this in Section 3.3.3 where we used Student-t errors in a regression. The model was

$$\begin{aligned} Y_i \mid \mu_i, X_i &\sim N(\mu_i,\ X_i^{-1}), \\ d\sigma^2 X_i &\sim \chi^2_d. \end{aligned}$$

Here μ_i corresponds to λ_i in (4.4) and X corresponds to θ in (4.4). The mixing density is that of a scaled χ^2 distribution and $f_y(y_i; \theta, \lambda_i)$ in (4.4) corresponds to $\phi(X_i^{1/2}[y_i - \mu_i])$ where $\phi()$ is the standard normal pdf.

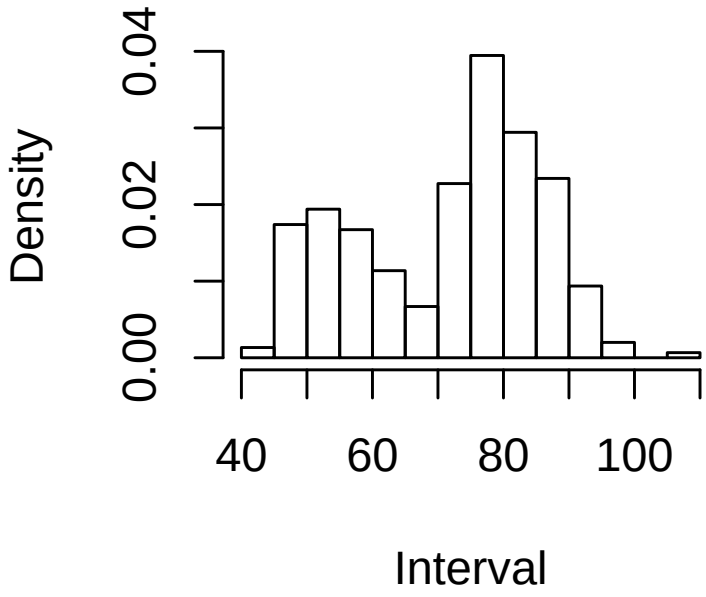
MAS8303: Modern Bayesian Inference

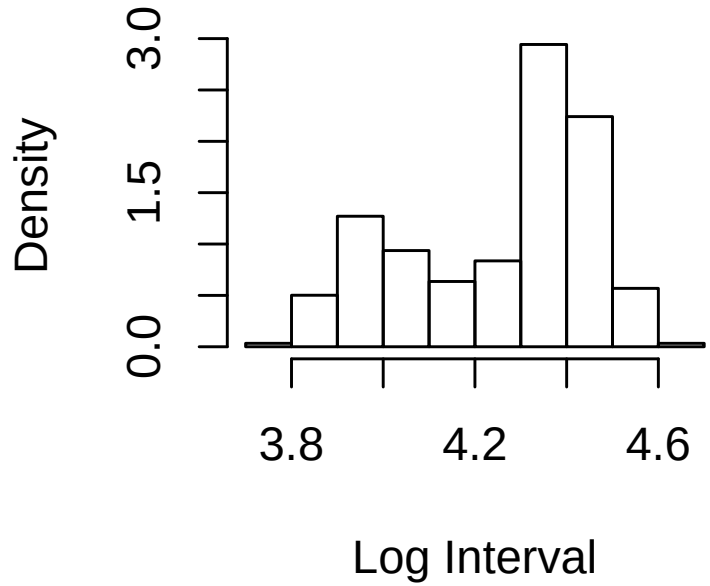
Malcolm Farrow

November 2010

Lecture 4.2: Mixture Examples: "Old Faithful"

Lecture 4.2: Mixture Examples: "Old Faithful"





Lecture 4.2: Mixture Examples: Normal mixture

Let us try using a two-component normal mixture model for the log intervals. So

$$
\begin{aligned}
\Pr(c_i = 1) &= \pi \\
\Pr(c_i = 2) &= 1 - \pi \\
\pi &\sim \mathrm{Beta}(a_\pi, b_\pi) \\
y_i \mid \mu_j, \tau_j, c_i = j &\sim N(\mu_j,\ \tau_j^{-1}) \\
\mu_j \mid \mu_0 &\sim N(\mu_0 + \delta_j,\ \tau_\mu^{-1}) \\
\mu_0 &\sim N(M_\mu,\ V_\mu) \\
\tau_j &\sim \mathrm{Ga}(a_\tau, b_\tau)
\end{aligned}
$$

Lecture 4.2: Mixture Examples: Normal mixture

Notice that we have given $\mu_1,\ \mu_2$ a "hierarchical prior." Each depends on μ_0 which then has a prior of its own. In order to avoid label switching we can impose the restriction $\mu_1 < \mu_2$. We also push the conditional prior means of μ_1, μ_2 apart by making them $\mu_0 + \delta_1$ and $\mu_0 + \delta_2$ respectively, where $\delta_1 = -\delta$ and $\delta_2 = \delta$.

Lecture 4.2: Mixture Examples: Normal mixture

We could also use a hierarchical prior for τ_1 and τ_2 although this is not quite as straightforward with gamma distributions as it is with normal distributions. I have just given them independent priors here. There is no need to impose an order constraint on $\tau_1,\ \tau_2$.

Lecture 4.2: Mixture Examples: Normal mixture

The specification of the prior is completed by giving numerical values to $a_\pi, b_\pi, a_\tau, b_\tau, M_\mu, V_\mu, \tau_\mu$. We will use the following values.

$$a_\pi = 4, \quad b_\pi = 4, \quad a_\tau = 4, \quad b_\tau = 0.04,$$

$$M_\mu = 4.0 \approx \log(60), \quad V_\mu = 0.30 \approx (\log(3)/2)^2,$$

$$\tau_\mu = 3.3 \approx (\log(3)/2)^{-2}, \quad \delta = 0.2.$$

Lecture 4.2: Mixture Examples: Normal mixture

```
model faithnorm

{for (i in 1:n)
   {c[i]~dcat(q[])
    y[i]~dnorm(mu[c[i]],tau[c[i]])
    }

 for (j in 1:2)
   {tau[j]~dgamma(4,0.04)
    }
```

Lecture 4.2: Mixture Examples: Normal mixture

```
mumean[1]<-mu0-0.2
mumean[2]<-mu0+0.2
mu[1]~dnorm(mumean[1],3.3) I(,mu[2]) # This imposes the
mu[2]~dnorm(mumean[2],3.3) I(mu[1],) # order constraint.

mu0~dgamma(4.0,p.mu)
p.mu<-1/0.3

pi~dbeta(3,3)
q[1]<-pi
q[2]<-1-pi
}
```

Lecture 4.2: Mixture Examples: Gamma mixture

As an alternative to the normal mixture for the log intervals, which is, of course, equivalent to a lognormal mixture for the intervals, we could try a gamma mixture for the intervals themselves.

$$
\begin{aligned}
\Pr(c_i = 1) &= \pi \\
\Pr(c_i = 2) &= 1 - \pi \\
\pi &\sim \mathrm{Beta}(a_\pi, b_\pi) \\
t_i \mid \alpha_j, \beta_j, c_i = j &\sim \mathrm{Ga}(\alpha_j,\ \beta_j) \\
\beta_j &= \alpha_j / \lambda_j \\
\lambda_j &= \exp(\mu_j) \\
\mu_j \mid \mu_0 &\sim N(\mu_0 + \delta_j,\ \tau_\mu^{-1}) \\
\mu_0 &\sim N(M_\mu,\ V_\mu) \\
\alpha_j &\sim \mathrm{Ga}(a_\alpha, b_\alpha)
\end{aligned}
$$

Lecture 4.2: Mixture Examples: Gamma mixture

Since the mean of a Ga(α_j, β_j) distribution is α_j/β_j and we set $\beta_j = \alpha_j/\lambda_j$, the mean interval, in component j, is λ_j. We then treat $\mu_j = \log(\lambda_j)$ in the same way as we treated μ_j in the lognormal mixture. Of course the log of the mean is not the same as the mean of the logs but, in this case, this difference has little effect. (To avoid this slight discrepancy we would have to make λ_j the median rather than the mean but this is not convenient with a gamma distribution).

Lecture 4.2: Mixture Examples: Gamma mixture

I have not used a hierarchical prior for α_1, α_2. I have just given them independent priors here. There is no need to impose an order constraint on $\alpha_1,\ \alpha_2$.

Lecture 4.2: Mixture Examples: Gamma mixture

We will use the following values to complete the prior specification.

$$a_\pi = 1, \quad b_\pi = 1, \quad a_\alpha = 3, \quad b_\alpha = 0.1,$$

$$M_\mu = 4.0 \approx \log(60), \quad V_\mu = 0.30 \approx (\log(3)/2)^2,$$

$$\tau_\mu = 3.3 \approx (\log(3)/2)^{-2}, \quad \delta = 0.2.$$

Lecture 4.2: Mixture Examples: Gamma mixture

```
model faithgamma

{for (i in 1:n)
   {c[i]~dcat(q[])
    t[i]~dgamma(alpha[c[i]],beta[c[i]])
    }

 for (j in 1:2)
   {alpha[j]~dgamma(3,0.1)
    beta[j]<-alpha[j]/lambda[j]
    lambda[j]<-exp(mu[j])
    }
```

Lecture 4.2: Mixture Examples: Gamma mixture

```
mumean[1]<-mu0-0.2
mumean[2]<-mu0+0.2
mu[1]~dnorm(mumean[1],3.3) I(,mu[2]) # This imposes the
mu[2]~dnorm(mumean[2],3.3) I(mu[1],) # order constraint.

mu0~dnorm(4.0,p.mu)
p.mu<-1/0.3

pi~dbeta(1,1)
q[1]<-pi
q[2]<-1-pi
}
```

Lecture 4.2: Mixture Examples: Headways

Time gaps, or "headways", between vehicles passing along a road. The idea is that headways fall naturally into one of two sub-populations:

1. Headways where the following vehicle is not impeded by the vehicle in front.
2. "Congested" headways where the following vehicle is impeded by the vehicle in front.

Lecture 4.2: Mixture Examples: Headways

1: non-congested headways : Exponential distribution, that is $\text{Ga}(1, \beta_1)$.

2: Congested headways : $\text{Ga}(\alpha_2, \beta_2)$ distribution with $\alpha_2 > 1$.

Lecture 4.2: Mixture Examples: Headways

- The constraint that $\alpha_2 > 1$ is imposed by letting $\alpha_2 = 1 + A$ where $A \sim \mathrm{Ga}(a_A, b_A)$. We have $a_A = 2$ and $b_A = 8$.
- The mean headway in component 1 is $\mu_1 = \beta_1^{-1}$.
- The mean headway in component 2 is $\mu_2 = \alpha_2/\beta_2$.
- We set $\beta_2 = \alpha_2 B$ where $B \sim \mathrm{Ga}(a_B, b_B)$. We have $a_B = 1$ and $b_B = 2$. Thus $\mu_2 = B^{-1}$.
- By imposing the constraint $B > \beta_1$ we ensure that $\mu_1 > \mu_2$.

Lecture 4.2: Mixture Examples: Headways

```
model headway;

{

 for (i in 1:N)
     {c[i]~dcat(q[])
      t[i]~dgamma(alpha[c[i]],beta[c[i]])
      }
```

Lecture 4.2: Mixture Examples: Headways

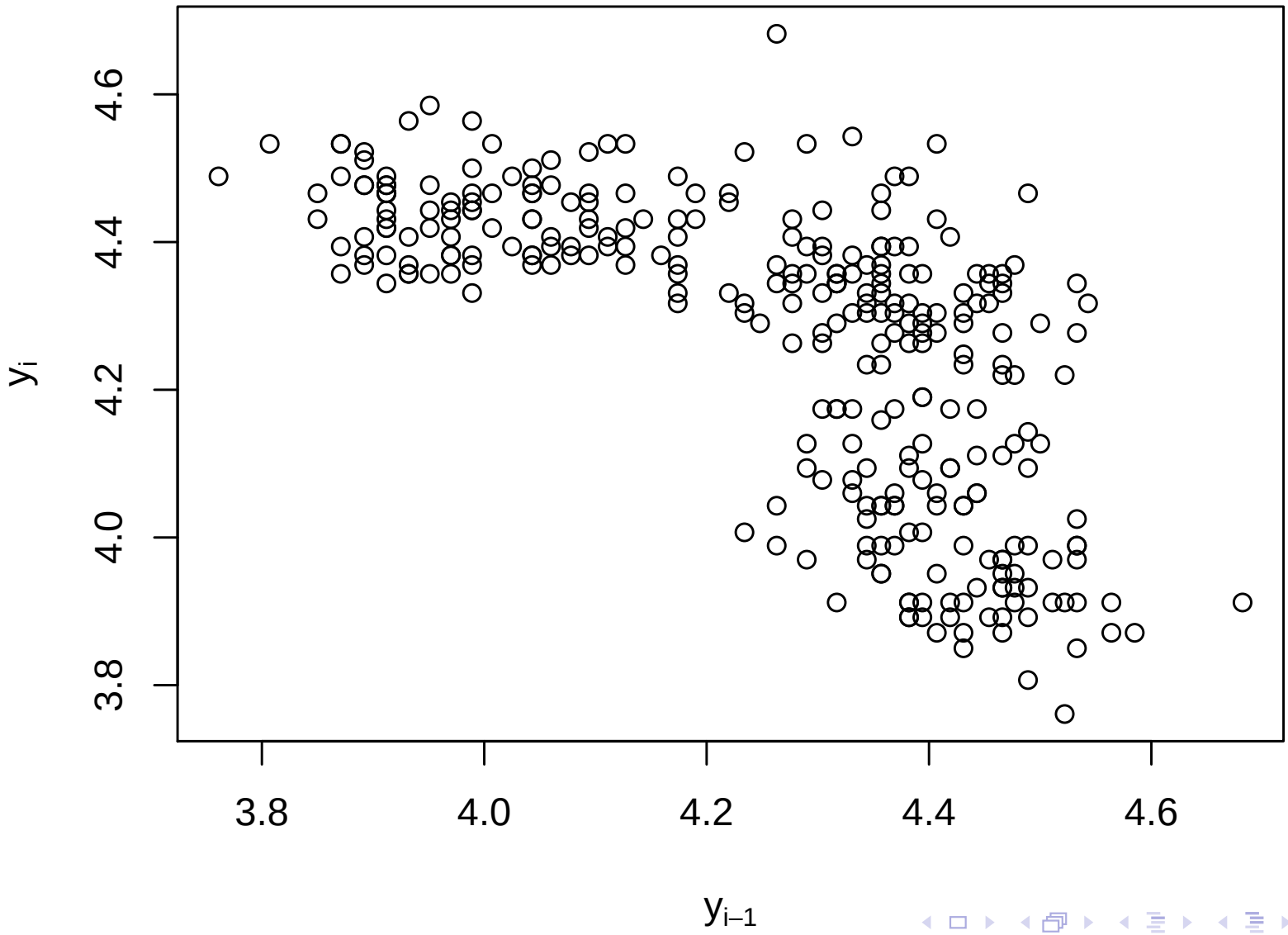
```
alpha[1]<-1
alpha[2]<-1+aa
aa~dgamma(1,0.5)
beta[1]~dgamma(2,8) I(,bb)
beta[2]<-alpha[2]*bb
bb~dgamma(1,2) I(beta[1],)
pi~dbeta(1,2)
q[1]<-pi
q[2]<-1-pi

mu[1]<-1/beta[1]
mu[2]<-1/bb

}
```

MAS8303: Modern Bayesian Inference

Malcolm Farrow

November 2010

Lecture 4.3: Hidden Markov Models



Lecture 4.3: Hidden Markov Models

We could model the sequence $c_1, \ldots, c_n$ using a two-state Markov chain with the folowing transition matrix, where $q_{j,k} = \Pr(c_i = j \mid c_{i-1} = k)$.

$$\begin{pmatrix} q_{1,1} & q_{1,2} \\ q_{2,1} & q_{2,2} \end{pmatrix} = \begin{pmatrix} 0 & \pi \\ 1 & 1-\pi \end{pmatrix}. \quad (4.5)$$

Lecture 4.3: Hidden Markov Models

Of course, before we saw the data we would not know about this pattern so it could be argued that we should use a more general model in which we allow $q_{1,1} > 0$. In this case we would have

$$\begin{pmatrix} q_{1,1} & q_{1,2} \\ q_{2,1} & q_{2,2} \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 \\ 1-\pi_1 & 1-\pi_2 \end{pmatrix}. \quad (4.6)$$

Lecture 4.3: Hidden Markov Models

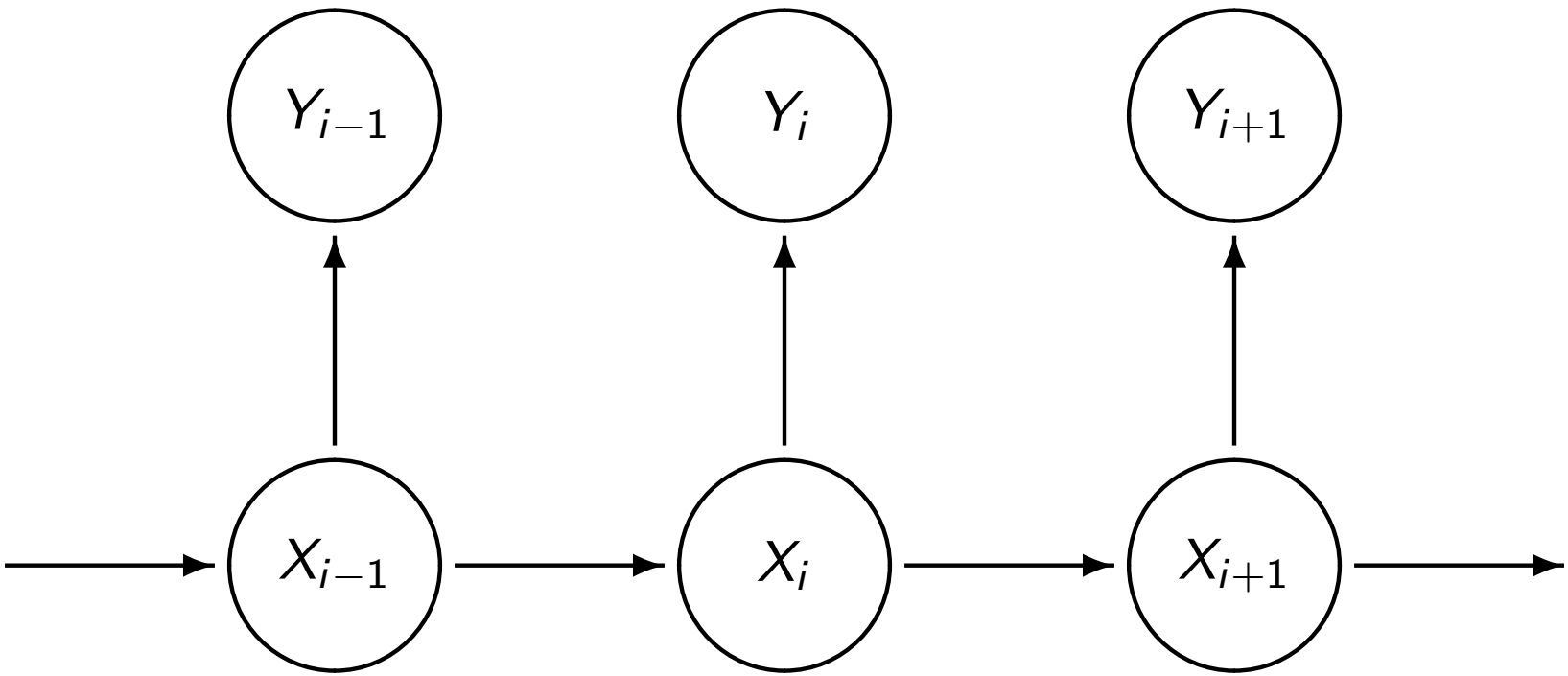
Examples of *hidden Markov models* or HMM:

- Time series,
- DNA sequences,
- Linguistics,
- etc.

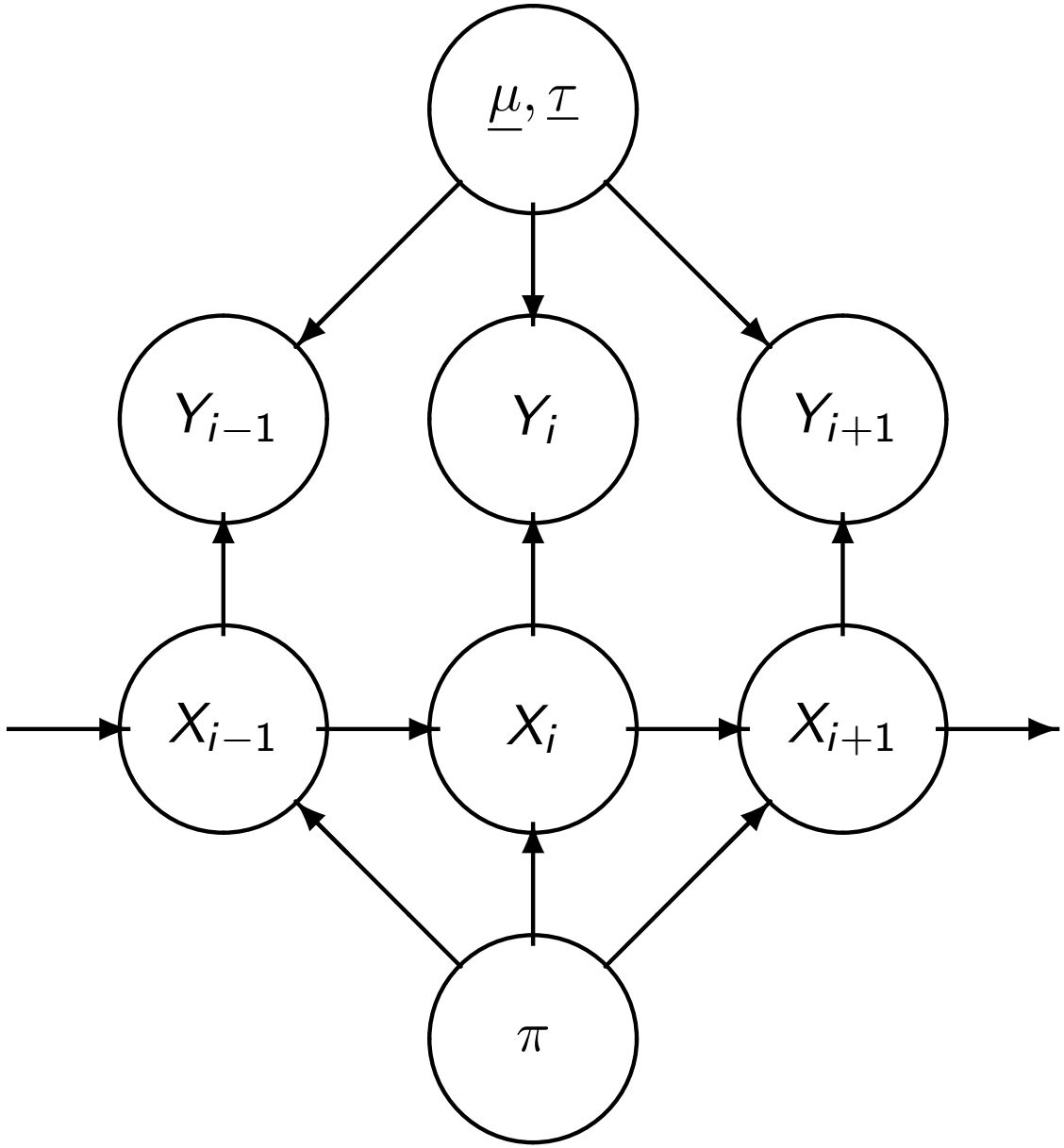
Lecture 4.3: Hidden Markov Models

In general, in a HMM, we have a sequence of (possibly vector) observations $\ldots y_{i-1}, y_i, y_{i+1} \ldots$ where the distribution of y_i depends on the value of an unobserved (i.e. *latent*) (possibly vector) variable x_i and the sequence $\ldots x_{i-1}, x_i, x_{i+1}, \ldots$ forms a Markov chain.

Lecture 4.3: Hidden Markov Models: Figure 4.6



Lecture 4.3: Hidden Markov Models: Figure 4.7



Lecture 4.3: Hidden Markov Models: Old Faithful

Stationary distribution of the Markov chain, in this case, is

$$\begin{aligned} \Pr(c_1 = 1) &= \frac{\pi}{1-\pi}, \\ \Pr(c_1 = 2) &= \frac{1}{1-\pi}. \end{aligned}$$

Lecture 4.3: Hidden Markov Models: Old Faithful

```
model faithnormhmm

{p0[1]<-0.5
 p0[2]<-0.5
 cc[1]~dcat(p0[])

 for (i in 2:30)
   {cc[i]~dcat(q[,cc[i-1]])    # This is the
    }                          # "burn-in" section.

 c[1]~dcat(q[,cc[30]])       # This is for
                             # the initial state.
```

Lecture 4.3: Hidden Markov Models: Old Faithful

```
for (i in 2:n)
  {c[i]~dcat(q[,c[i-1]])
   }

for (i in 1:n)
  {y[i]~dnorm(mu[c[i]],tau[c[i]])
   }

for (j in 1:2)
  {tau[j]~dgamma(4,0.04)
   }
```

Lecture 4.3: Hidden Markov Models: Old Faithful

```
mumean[1]<-mu0-0.2
mumean[2]<-mu0+0.2
mu[1]~dnorm(mumean[1],3.3) I(,mu[2]) # This imposes the
mu[2]~dnorm(mumean[2],3.3) I(mu[1],) # order constraint.

mu0~dnorm(4.0,p.mu)
p.mu<-1/0.3

q[1,2]<-pi
pi~dbeta(1,1)
q[2,2]<-1-q[1,2]
q[1,1]<-0.0
q[2,1]<-1-q[1,1]

}
```

Lecture 4.3: Hidden Markov Models: Road traffic headways

The stationary distribution of the Markov chain in this case has

$$\begin{aligned} \Pr(c_1 = 1) &= \frac{\pi_2}{1 + \pi_2 - \pi_1}, \\ \Pr(c_1 = 2) &= \frac{1 - \pi_1}{1 + \pi_2 - \pi_1}. \end{aligned}$$

Lecture 4.3: Hidden Markov Models: Road traffic headways

```
model headway

{p0[1]<-0.5
 p0[2]<-0.5
 cc[1]~dcat(p0[])
 for (i in 2:30)
     {cc[i]~dcat(q[,cc[i-1]])    # This is the
      }                          # "burn-in" section.

 c[1]~dcat(q[,cc[30]])
 t[1]~dgamma(alpha[c[1]],beta[c[1]])
```

Lecture 4.3: Hidden Markov Models: Road traffic headways

```
for (i in 2:N)
    {c[i]~dcat(q[,c[i-1]])
     t[i]~dgamma(alpha[c[i]],beta[c[i]])
     }
```

Lecture 4.3: Hidden Markov Models: Road traffic headways

```
alpha[1]<-1
alpha[2]<-1+aa
aa~dgamma(1,0.5)
beta[1]~dgamma(2,8) I(,bb)
beta[2]<-alpha[2]*bb
bb~dgamma(1,2) I(beta[1],)
pi[1]~dbeta(1,2)
pi[2]~dbeta(1,2)
q[1,1]<-pi[1]
q[1,2]<-pi[2]
q[2,1]<-1-pi[1]
q[2,2]<-1-pi[2]
```

Lecture 4.3: Hidden Markov Models: Road traffic headways

```
 lrr<-log(pi[1]/pi[2])
 pos<-step(lrr)
 mu[1]<-1/beta[1]
 mu[2]<-1/bb
}
```