SC7/SM6 Bayes Methods HT19

Lecturer: Geoff Nicholls

Lecture 12: Reversible Jump MCMC II.

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/BayesMethods/

and via the MSc weblearn pages.
Reversible Jump MCMC*

Consider now a set of models \( p(y|\theta, m) \) and \( \pi(\theta|m), \theta \in \Omega_m \), with model prior \( \pi(m), m = 1, \ldots, M \). We are targeting

\[
\pi(\theta, m|y) \propto p(y|\theta, m)\pi(\theta|m)\pi(m).
\]

The MCMC state is \( X_t = (\theta, m) \). In the following we know \( m \) if we know \( \theta \) (\( m \) might be the dimension of \( \theta \) for example) so we could drop the \( m \) reference if we wish.

Let \( \rho_{m,m'} \) be a matrix of proposal probabilities: the probability to propose a move to model \( m' \) given the current state is \( m \).

Suppose model $m'$ has one more parameter than $m$, for example,

$$\theta|m = (\theta_1, ..., \theta_d)$$

and

$$\theta'|m' = (\theta'_1, ..., \theta'_d, \theta'_{d+1}),$$

and we have created a move $(\theta', u') = \psi(\theta, u)$ with

$$\psi(\theta, u) = (\psi_1(\theta, u), \psi_2(\theta, u))$$

that generates $\theta'$ from $\theta$. This creates a flow from $\theta \rightarrow \theta'$ and we need to balance it.

To set the mapping $\psi$ up as an differentiable involution extend the space $U$ of the proposal $g(u)$ with another state, $\{\emptyset\}$ so that $\Omega_{m'} \times \{\emptyset\}$ and $\Omega_{m'} \times U$ have equal dimension $d + 1$.

Extend the definitions of $\psi_1$ and $\psi_2$ so that $\theta = \psi_1(\theta', \emptyset)$ and $u = \psi_2(\theta', \emptyset)$ and hence $(\theta, u) = \psi(\theta', \emptyset)$.

We extend $g(u)$ to $G_m(m', u) = \rho_{m', m}g(u)$ when the dimension of $\Omega_m$ is greater than that of $\Omega_m'$ and $G_{m'}(m, \emptyset) = \rho_{m', m}$. 
Example Proposal $d \rightarrow d + 1$: selected with probability $\rho_{m,m'}$; simulate $u \sim g(u)$ and set

$$\theta'_{1:d+1} = \psi_1(\theta, u)$$

where $\theta'_{1:d} = \theta_{1:d}$ and

$$\theta'_{d+1} = \theta'_{d+1}(\theta_{1:d}, u)$$

for the last component. For simplicity assume $\theta'_{d+1} = \theta'_{d+1}(u)$.

Proposal $(d + 1 \rightarrow d)$: selected with probability $\rho_{m',m}$; set $\theta_i = \theta_i', i = 1, ..., d$ (i.e., delete $\theta'_{d+1}$). In this move $u' = \emptyset$ and

$$\psi(\theta'_{1:d+1}, u') = (\theta_{1:d}, u),$$

where $u$ solves $\theta'_d = \theta'_d(u)$ (it is the $u$ that “takes us back”).
The Jacobian is
\[
\left| \frac{\partial \psi(\theta, u)}{\partial \psi(\theta', u')} \right| = \left| \frac{\partial \theta'}{\partial (\theta, u)} \right| = \left| \frac{\partial \theta'_{d+1}}{\partial u} \right|.
\]

Dimensions are matched as \( \dim(\theta) + \dim(u) = \dim(\theta') \). The acceptance probability for proposal \( d \rightarrow d + 1 \) is
\[
\alpha(\theta'|\theta) = \min \left\{ 1, \frac{\pi(\theta', m'|y) G_{m'}(m, u)}{\pi(\theta, m|y) G_{m'}(m, u)} \left| \frac{\partial \theta'}{\partial (\theta, u)} \right| \right\}
\]
\[
= \min \left\{ 1, \frac{\pi(\theta', m'|y) \rho_{m',m}}{\pi(\theta, m|y) \rho_{m,m'} q(\theta'_{d+1})} \right\}.
\]

where
\[
q(\theta'_{d+1}) = g(u) \left| \frac{\partial \theta'_{d+1}}{\partial u} \right|^{-1}
\]
is just the proposal distribution for \( \theta'_{d+1} \) (if the proposals depended on \( \theta \) then this would be \( q(\theta'_{d+1}|\theta) \).
The acceptance probability for the reverse move, \( d + 1 \rightarrow d \) from \( \theta' = (\theta'_1, ..., \theta'_d, \theta'_{d+1}) \) to \( \theta = (\theta'_1, ..., \theta'_d) \) is just the inverse
\[
\alpha(\theta|\theta') = \min \left\{ 1, \frac{\pi(\theta, m|y)\rho_{m,m'}q(\theta_{d+1})}{\pi(\theta', m'|y)\rho_{m',m}} \right\}.
\]

The setup described above is a special case, jumping one dimension, with a simple proposal scheme. However, the framework generalises a great deal.

For example if the state is a set \( x = \{x_1, ..., x_m\} \) we might choose the element to delete at random. That probability will enter detailed balance as the choice of update involves the choice of \( m' \) and \( i \). The acceptance probability for adding an element \((x, m) \rightarrow (x', m + 1)\) with \( x' = x \cup \{x_{m+1}\} \) is
\[
\alpha(x'|x) = \min \left\{ 1, \frac{\pi(x', m + 1|y)\rho_{m+1,m} \times \frac{1}{m+1}}{\pi(x, m|y)\rho_{m,m+1}q(x_{m+1})} \right\}.
\]
RJMCMC - simple example

Let \( X = 1/2 \) with probability 1/3, else \( X \sim 2xI_{0<x<1}. \)

In this example we have a mixture of two models of dimension zero and one: Model \( M = 1 \) has state space \( X|M = 1 \in \{1/2\} \) (ie, a point); Model \( M = 2 \) has state space \( X|M = 2 \in [0, 1] \).

In terms of the joint (value,model) pair the target pmf/pdf is

\[
\pi(x, m) = \pi(x|m)\pi(m)
\]

with \( \pi(m = 1) = 1/3, \pi(m = 2) = 2/3, \pi(x|m = 1) = I_{x = 1/2}, \pi(x|m = 2) = 2x. \) We would like to give a RJ MCMC algorithm targeting this distribution.

\( \dagger \)The CDF is \( F_X(x) = \Pr(X \leq x) \) with \( F_X(x) = \frac{2}{3}x^2 + \frac{1}{3}I_{x \geq 1/2} \)
**RJMCMC algorithm targeting** \((X, M) \sim \pi(x, m)\):

MCMC state: \((X_t, M_t) = (x, m)\). First the proposal rules:

(increase dimension) if \(m = 1\) propose \(m' = 2\) with probability \(\rho_{1,2} = 1\). Propose \(x' \sim q(x')\) (recall \(x'\) is a scalar rv in \([0, 1]\)).

(decrease dimension) if \(m = 2\) propose \(m' = 1\) with probability \(\rho_{2,1} = 1\) and set \(x' = 1/2\).

We can choose \(q(x')\) to be anything we like that is irreducible. Just to prove this all works I use

\[
q(x') = \text{Beta}(x'; \alpha = 1/2, \beta = 1/2)
\]

ie something dramatically different from the density \(2x\) we expect when \(m = 2\).
Acceptance probabilities:
If \((x, m) = (1/2, 1)\) and we propose \((x', m' = 2)\) (ie, increase dimension), the acceptance probability is

\[
\alpha(x', m'|x, m) = \min \left\{ 1, \frac{\pi(x'|m')\pi(m')\rho_{m',m}}{\pi(x|m)\pi(m)\rho_{m,m'}q(x'|m')} \right\}.
\]

Substituting in our values the AP is

\[
\alpha(x', m'|x, m) = \min \left\{ 1, \frac{4x'/3}{\text{Beta}(x'; \alpha, \beta)/3} \right\}.
\]

If \((x, m) = (x, 2)\) and we propose \((x' = 1/2, m' = 1)\) (ie, decrease dimension), the acceptance probability is

\[
\alpha(x', m'|x, m) = \min \left\{ 1, \frac{\pi(x'|m')\pi(m')\rho_{m',m}q(x|m')} {\pi(x|m)\pi(m)\rho_{m,m'}} \right\}.
\]
Substituting in our values the AP is
\[
\alpha(x', m'|x, m) = \min \left\{ 1, \frac{\text{Beta}(x; \alpha, \beta)/3}{4x/3} \right\}.
\]

Iteration: we generate our chain \((X_t, M_t)\) iterating proposals and acceptance steps using the formula above. If the model is \(m = 1\) (so the state is \(x = 1/2\)) we propose to jump to model \(m' = 2\) and a new state \(x' \in [0, 1]\), and vis versa.
Remark 1: the dimension of the proposal dbn matches the change in dimension in the target - in terms of our original notation we propose a switch from $m = 1$ to $m' = 2$, simulate $u \sim g(u)$ and set $(x', u') = \psi(x, u)$. Here $u \sim \text{Beta}(\cdot; \alpha, \beta)$, $u' = \emptyset$ and the “transformation” is just $(x', u') = (u, \emptyset)$ so the Jacobian $|\partial x'/\partial u|$ is one. Also $\mathcal{U} = [0, 1]$, $\Omega_1 = \{1/2\}$ and $\Omega_2 = [0, 1]$ so
\[
\dim(\mathcal{U} \times \Omega_1) = \dim(\Omega_2 \times \{\emptyset\})
\]
since both dimensions equal one.

Remark 2: we could if we wished mix in a fixed-dimension update (ie set $\rho_{2,1} = \rho_{2,2} = 1/2$ so we have two options if $m = 2$). In this fixed dimension update we target $\pi(x|m = 2)$ using our standard MCMC tools.
Summarising the posterior from RJ-MCMC output

RJ MCMC is a Monte Carlo method useful for Bayesian model selection and model averaging. If we can sample the joint posterior for model \( m \) and parameter \( \theta \)

\[
\theta^{(t)}, m^{(t)} \sim \pi(\theta, m|y)
\]

we can carry out model averaging and model selection.

**Model Choice:** Since \( m^{(t)} \sim \pi(m|y) \) (i.e., marginally), the maximum a posteriori model (the MAP)

\[
m_{\text{MAP}} = \arg \max_{m=1, \ldots, M} \pi(m|y)
\]

can be estimated by the mode of \( \{m^{(t)}\}_{t=1}^T \). The MAP model minimises the risk for the 0-1 loss function \( L(m^*, m) = \mathbb{I}_{m=m^*} \).
When the number of models is very large, we can summarise the uncertainty over models using an HPD credible set $C$ over models. This set satisfies

$$\sum_{m \in C} \pi(m|y) = 1 - \alpha^*$$

and

$$\pi(m|y) \geq \pi(m'|y)$$

for all pairs $m \in C$ and $m' \in C^c$. The HPD set minimises the risk for the loss to cover the true model $m_0$,

$$L(m_0, C) = c \mathbb{I}_{m_0 \notin C} + \text{card}(C),$$

where $c$ depends on $\alpha$ (see *The Bayesian Choice* Section 5.5.3).

*or as close as we can manage given $m$ is discrete.*
Parameter estimation: Since $\theta^{(t)} \sim \pi(\theta|y)$, the model averaged posterior expectation $E_{\Theta|Y=y}(h(\Theta)) = E_{\Theta,M|Y=y}(h(\Theta))$ can be estimated by the mean of $\{h(\theta^{(t)})\}_{t=1}^{T}$. For prediction or goodness of fit checking, the posterior predictive distribution

$$p(y'|y) = \sum_m \int p(y'|\theta, m)\pi(\theta, m|y)d\theta$$

can be simulated via $y^{(t)} \sim p(\cdot|\theta^{(t)}, m^{(t)})$ or estimated via

$$\overline{p(y'|y)} = \frac{1}{T} \sum_{t=1}^{T} p(y'|\theta^{(t)}, m^{(t)}).$$

When carrying out model selection, the posterior predictive distribution conditioned on the selected model

$$p(y'|y, \hat{m}) = \int p(y'|\theta, \hat{m})\pi(\theta|y, \hat{m})d\theta$$

can be simulated and estimated in a similar way and plotted over the data (for a quick visual check, as a test needs reserved data).
RJ MCMC and fitting mixture models

The Galaxy radial velocity data are shown in the figure below. It is natural to model this via a mixture of normals. However we do not know the number of components in the mixture.
Likelihood: Our data \( y_i \in \mathbb{R}, i = 1, 2, \ldots, n \) are independent samples from a mixture model with \( m \) components \( N(\mu_j^{(m)}, \sigma_j^{(m)}^2) \), and mixture weights \( w_j^{(m)}, j = 1, 2, \ldots, m, w_j > 0, \sum_{j=1}^m w_j = 1 \).

Given \( m \in 1, 2, 3, \ldots \) the sets of mixture parameters are

\[
\mu^{(m)} = \{\mu_1^{(m)}, \ldots, \mu_m^{(m)}\}, \quad \sigma^{(m)} = \{\sigma_1^{(m)}, \ldots, \sigma_m^{(m)}\}, \quad w^{(m)} = \{w_1^{(m)}, \ldots, w_m^{(m)}\}.
\]

The observation model for the iid \( y_i, i = 1, 2, \ldots, n \) is the mixture

\[
(y_i | \mu^{(m)}, \sigma^{(m)}, w^{(m)}) \sim \sum_{j=1}^m w_j^{(m)} N(y_i; \mu_j^{(m)}, \sigma_j^{(m)}^2).
\]

The likelihood is therefore

\[
L(\mu^{(m)}, \sigma^{(m)}, w^{(m)}, m; y) = \prod_i \left[ \sum_{j=1}^m w_j^{(m)} N(y_i; \mu_j^{(m)}, \sigma_j^{(m)}^2) \right].
\]
Priors: We take as our priors

\[ w^{(m)} \sim \text{Dirichlet}(\alpha 1_m) \]

with \( 1_m \) a vector of \( m \) ones, \( \alpha = 1 \) (\( w^{(m)} \) uniform, sum to one),

\[ \mu_j^{(m)} \sim N(20, 10), \quad \text{iid for } j = 1, 2, \ldots, m, \]

which of course covers the data (covers \([0, 40]\) at \( 2\sigma \) - I assume the scale of the response is known), and

\[ \sigma_j^{(m)} \sim \text{Gamma}(1.5, 0.5), \quad \text{iid for } j = 1, 2, \ldots, m, \]

again informed by the scale: mean equals 3; shape 1.5 rules out very dense clusters at small \( \sigma \); small rate gives heavy tail, standard deviation about 2.5. For a model prior I take \( m \sim \text{Poisson}(\lambda|m > 0) \) with \( \lambda = 10 \), which is centred at 10, and tails off above about 20 clusters.
**Posterior:** The posterior for the model and parameters

$$\theta^{(m)} = (\mu^{(m)}, \sigma^{(m)}, w^{(m)})$$

is, for \( m = 1, 2, 3, \ldots \),

$$\pi(\theta^{(m)}, m|y) \propto L(\mu^{(m)}, \sigma^{(m)}, w^{(m)}, m; y)$$
$$\times \text{Dirichlet}(w^{(m)}; \alpha 1_m)$$
$$\times \prod_{j=1}^{m} N(\mu_j^{(m)}; 20, 10) \text{Gamma}(\sigma_j^{(m)}; 1.5, 0.5)$$
$$\times \text{Poisson}(m; \lambda)$$

Note that if \( m, m' \) are both greater than zero then

$$\text{Poisson}(m; \lambda|m > 0) \propto \text{Poisson}(m; \lambda),$$

so the condition has disappeared.
RJ MCMC algorithm fitting a normal mixture with an unknown number of components to the Galaxy Velocity data

Suppose the state is $X_t = (\mu, \sigma, w, m)$ with $\mu = (\mu_1, \ldots, \mu_m)$ etc. To get irreducibility we again need fixed dimension moves (3 of these) and variable dimension moves (2 of these).

Step 1. Choose an update UAR, $\text{move} \sim U\{1, 2, \ldots, 5\}$.

Step 2I. If $\text{move} = 1$ add a component (increase state dimension by three). Set $m' = m + 1$.

Step 2Ia Simulate $\mu'_{m+1}, \sigma'_{m+1} \sim q_{\mu\sigma}(\mu'_{m+1}, \sigma'_{m+1})$. We will take $q_{\mu\sigma}$ to be the Normal-Gamma prior above. Set $\mu' = (\mu, \mu'_{m+1})$ and $\sigma' = (\sigma, \sigma'_{m+1})$. 
Step 2Ib Now simulate $w'$. We have to make sure $\sum_j w'_j = 1$ (still). Choose a weight $j \sim U\{1, 2, \ldots, m\}$ to “split”. Simulate $w'_{m+1} \sim U(0, w_j)$ and for $k = 1, 2, \ldots, m + 1$ set

$$w'_k = \begin{cases} w_k & k = 1, \ldots, m, k \neq j \\ w_k - w'_{m+1} & k = j \\ w'_{m+1} & k = m + 1 \end{cases}$$

The probability to propose $m'$ given $m$ is $\rho_{m,m'} = 1/5$. The probability to propose $(\mu', \sigma', w')$ given $(\mu, \sigma, w)$ is

$$q(\mu', \sigma', w'|\mu, \sigma, w) = q_{\mu\sigma}(\mu'_{m+1}, \sigma'_{m+1}) \times \frac{1}{m} \times \frac{1}{w_j}.$$  

In the reverse move we will pick a component $i$ of the mixture at random, delete it and add its weight to a randomly chosen component $j$ out of the remainder. The probability to propose
this reverse move back from \((\mu', \sigma', w')\) to \((\mu, \sigma, w)\) is just

\[
p(i, j) = \frac{1}{m(m + 1)},
\]

(given \(m, m'\) already decided) since we must choose the two components involved in the update.

**Step 3I.** Accept the proposal \((\mu', \sigma', w', m')\) with probability

\[
\alpha^+ = \alpha(\mu', \sigma', w', m'|\mu, \sigma, w, m)
\]

where

\[
\alpha^+ = \min \left\{ 1, \frac{\pi(\mu', \sigma', w', m'|y)p(i, j)}{\pi(\mu, \sigma, w, m|y)q(\mu', \sigma', w'|\mu, \sigma, w)} \right\}
\]
Step 2D. If move = 2 delete a component (decrease state dimension by three). Set $m' = m - 1$ (if $m' = 0$, reject the move and set $X_{t+1} = X_t$).

Step 2Da Simulate $i \sim U\{1, 2, \ldots, m\}$. Set $\mu' = \mu_{-i}$, $\sigma' = \sigma_{-i}$.

Step 2Db To update $w$ (ensuring $w$ is still normalised), simulate $j \sim U\{1, 2, \ldots, m\} \setminus \{i\}$ and then (i) set $w' = w$, (ii) set $w'_j = w_j + w_i$, (iii) set $w' = w'_{-i}$.

The probability to propose $m'$ given $m$ is $\rho_{m,m'} = 1/5$ again. The probability to propose $(\mu', \sigma', w')$ given $(\mu, \sigma, w)$ is just

\[ p(i, j) = 1/m(m - 1), \]
and to propose the move back, \((\mu', \sigma', w') \rightarrow (\mu, \sigma, w)\), is

\[
q(\mu, \sigma, w|\mu', \sigma', w') = q_{\mu\sigma}(\mu'_m, \sigma'_m) \times \frac{1}{m - 1} \times \frac{1}{w_i + w_j}.
\]

Step 3D. Accept the proposal \((\mu', \sigma', w', m')\) with probability

\[
\alpha^- = \alpha(\mu', \sigma', w', m'|\mu, \sigma, w, m)
\]

where

\[
\alpha^- = \min \left\{ 1, \frac{\pi(\mu', \sigma', w', m'|y)q(\mu, \sigma, w|\mu', \sigma', w')}{\pi(\mu, \sigma, w, m|y)p(i, j)} \right\}
\]

We have additionally moves 3-5 which act on \(\mu, \sigma\) and \(w\) respectively in fixed dimension moves.
RJ-MCMC for a normal mixture
We illustrate the method on the Galaxy velocity distribution data. The R-code and further detail of the algorithm are available on the course website. We ran the code and generated samples \((\mu(t), \sigma(t), w(t), m(t))\), \(t = 1, 2, ..., T\) from the joint posterior distribution over the number of clusters and the cluster weights and parameters.

The plot above shows the posterior distribution over the number of components. 3-6 components is the number favored.
The plot shows traces for the log-prior, log-likelihood and number of components, (as the number of parameters vary, they are not easily plotted).
On the previous page, the bottom figure shows the sequence of sampled means $\mu^{(t)}, t = 1,\ldots,T$. The dimension of these vectors is not constant. A point at $(t, x)$ is colored by the index of $x$ in the vector $\mu^{(t)}$, i.e., the color is $i : \mu_{i}^{(t)} = x$. We can see label-switching in action. Stable mixture components (the horizontal bands) are often present in the state, but may appear in any position in the vector. The label of a mixture component is random (and uninteresting).

The top figure shows an estimate of posterior predictive distribution $p(y'\mid y)$ (black line) obtained by averaging $p(y'\mid \mu, \sigma, w, m)$ over the sampled states, at each point $y'$ on the $x$-axis, and the posterior predictive distribution $p(y'\mid y, m)$ conditioned on $m$ clusters (red is $m=3$, green is $m=4$, blue is $m=5$).

The underlying histogram in black is a histogram of the data, $y$. We expect the distribution of the data to match the posterior predictive distribution, and the fit seems reasonable.