SC7/SM6 Bayes Methods HT17

Lecturer: Geoff Nicholls

Lecture 11: Reversible Jump MCMC for model selection and model averaging

Notes and Problem sheets are available at

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

and via the MSc weblearn pages.
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 $m^{(t)} \sim \pi(m|y)$ (ie, 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 $\hat{m}$ say, of $\{m^{(t)}\}_{t=1}^{T}$. The MAP model is the choice minimising the risk for the 0-1 loss function $L(m^*, m) = \mathbb{I}_{m=m^*}$.

When the number of models is very large, we often summarise the uncertainty over models using an HPD credible set over models. This is just a standard level $\alpha$ HPD credible set $C$ for $m$ using the posterior $\pi(m|y)$. The HPD set minimises the risk for the loss

$$L(m^*, C) = c \mathbb{I}_{m^* \in C} + \text{card}(C),$$

where $c$ depends on $\alpha$ (see The Bayesian Choice Section 5.5.3).
Parameter estimation: Since $\theta^{(t)} \sim \pi(\theta|y)$, the model averaged posterior expectation $E_{\Theta|Y=y}(\Theta) = E_{\Theta,M|Y=y}(\Theta)$ can be estimated by the mean, $\hat{\theta}$ say, of $\{\theta^{(t)}\}_{t=1}^T$. The PE minimises the risk for the loss function

$$L(\theta^*, \theta) = |\theta^* - \theta|^2.$$  

If we want to do prediction, and for goodness of fit checking, the posterior predictive distribution

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

($L_m$ is the likelihood under model $m$) can be simulated or compared with the data. When carrying out model selection, the posterior predictive distribution conditioned on the selected model

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

can be plotted over the data (for a quick visual check, a test needs reserved data).
Example: variable selection in a normal linear model.

Consider variable selection for the normal linear model $Y \sim N(X\theta, \sigma^2)$, $\theta = (\theta_1, \ldots, \theta_p)$ with $X = [X_1, X_2, \ldots, X_p]$, $X_1 = 1$, $X_i = X_2^{i-1}$, $i = 2, 3, \ldots, p$.

Let $q = (q_1, \ldots, q_d)$ be a vector of covariate indices, with $0 \leq d \leq p$ and $1 \leq q_i \leq p$. Let $\theta(q) = (\theta_{q1}(q), \ldots, \theta_{qd}(q))$, so $\theta(q) \in \mathbb{R}^d$, and let $X(q) = X[\cdot, q]$. We allow $d = 0$ so $q = \phi$ (empty set), and define $X(\phi) = [0]$, a column of $n$ zeros in this case.

Here $q$ is a model index, with prior $\pi(q)$. The joint posterior is

$$
\pi(\theta(q), \sigma, q|y) \propto L(\theta(q), \sigma; y)\pi(\theta(q)|q)\pi(q),
$$

with $L(\theta(q), \sigma; y) = N(y, X(q)\theta(q), \sigma^2)$. 

**Priors** In this example we focus on the RJ MCMC and model summaries, so we will take priors as given. We scale the covariates and take $\beta_i \sim N(0, 3)$ and $\sigma \sim 1/\sigma$ in the parameter priors. The choice of model prior

$$
\pi(q) \propto \xi^d (1 - \xi)^{p-d}
$$

with say $\xi = c/p$ (with $c$ small, I use $c = 1$ below) is natural for variable selection in the usual regression context with large $p$. In polynomial regression and in dealing with interaction terms we might impose

$$
q \in \{\phi, (1), (1, 2), (1, 2, 3), ..., (1, 2, ..., p)\}
$$

which in our setting means all lower order interactions below the highest are automatically included.
RJ MCMC algorithm for normal linear regression

Suppose the state is $X_t = (\theta^{(q)}, q, \sigma)$. To get irreducibility over the full space

$$(\theta, q, \sigma) \in \Omega \quad \Omega = \bigcup_{q \subseteq 1:p} R^{\text{card}(q)} \times \{q\} \times R$$

we need fixed dimension moves and variable dimension moves. We may add extra fixed dimension moves to improve mixing.

Step 1. Choose an index $i \sim U\{1, 2, \ldots, p\}$. If $i \in q$ then $X_i$ is in the model and we propose to remove it. If $i \in \{1, 2, \ldots, p\} \setminus q$ then $X_i$ is not in the model and we propose to add it. Notice that the forward and backward model proposal probabilities are equal $\rho_{q, \tilde{q}} = \rho_{\tilde{q}, q} = 1/p$. 
If \( i \in \{1, 2, \ldots, p\} \setminus q \): Increase Dimension

Step 2I. Set \( \tilde{q} = (q, i) \). Simulate \( \tilde{\theta}^{(\tilde{q})} \sim q(\cdot) \) and set \( \tilde{\theta}^{(\tilde{q})} = (\theta^{(q)}, \tilde{\theta}^{(\tilde{q})}_i) \). We will take \( q(x) = N(x; 0, 3) \) to match the prior.

Step 3I. With probability \( \alpha^+ = \alpha(\tilde{\theta}^{(\tilde{q})}, \tilde{q}, \sigma | \theta^{(q)}, q, \sigma) \)

\[
\alpha^+ = \min \left\{ 1, \frac{\pi(\tilde{\theta}^{(\tilde{q})}, \tilde{q}, \sigma | y)}{\pi(\theta^{(q)}, q, \sigma | y) N(\tilde{\theta}^{(\tilde{q})}_i; 0, 3)} \right\}
\]

set \( X_{t+1} = (\tilde{\theta}^{(\tilde{q})}, \tilde{q}, \sigma) \) and otherwise \( X_{t+1} = X_t \).

\( \mathbb{P}(\Theta^* | S, y) \propto p(S) T(\Theta^* | S, y) \)
If $i \in q$: Decrease Dimension

Step 2D. Set $\tilde{q} = q \setminus \{i\}$ and set $\tilde{\theta}(\tilde{q}) = \theta(\tilde{q})$ (ie drop entry $\theta_i^{(q)}$ from $\theta^{(q)}$).

Step 3D. With probability $\alpha^- = \alpha(\tilde{\theta}(\tilde{q}), \tilde{q}, \sigma | \theta^{(q)}, q, \sigma)$

$$\alpha^- = \min \left\{ 1, \frac{\pi(\tilde{\theta}(\tilde{q}), \tilde{q}, \sigma | y) N(\theta_i^{(q)}; 0, 3)}{\pi(\theta^{(q)}, q, \sigma | y)} \right\}$$

set $X_{t+1} = (\tilde{\theta}(\tilde{q}), \tilde{q}, \sigma)$ and otherwise $X_{t+1} = X_t$.

We have additionally standard fixed dimension updates on $\theta^{(q)}$ for fixed model index $q$, and a standard update on $\sigma$ (the scaling update is convenient).
Summarising the Posterior over models and parameters for the NLM regression

We illustrate the method on data for average claims paid per policy for automobile insurance in New Brunswick in the years 1971-1980 (data from a Richard Lockhart ANOVA example).

The R-code and further detail of the algorithm are available on the course website. We ran the code and generated samples

\[
(\theta^{(q,t)}, q^{(t)}, \sigma^{(t)}) \sim \pi(\theta, q, \sigma | y) \quad t = 1, 2, ..., T.
\]
The traces show the model variation. The posterior predictive distribution is represented by plotting the function $v(x)\theta(q,t)$ against $x$ with $v = (1, x, x^2, x^3, x^4)$.

The data are black dots, the samples are grey, and the red curve $\hat{\mu}(x)$ say, is an estimate of the expectation of the posterior predictive distribution over $y' \sim p(y'|y)$ at $x$,

$$\mu(x) = E_{\theta,q,\sigma}(E_{Y'}(Y'|\theta,q,\sigma)|y)$$

which is just $\hat{\mu}(x) = v(x)\hat{\theta}$ with $\hat{\theta}$ the posterior mean averaged over models.
The posterior probabilities (as percentages) sorted by magnitude are

\[\begin{align*}
11110 & \quad 10110 & \quad 00010 & \quad 01010 & \quad 00110 & \quad 01000 & \quad 10010 & \quad 36.2 \\
01110 & \quad 10111 & \quad 11111 & \quad 00011 & \quad 11011 & \quad 10011 & \quad 11010 & \quad 1.9 \\
01011 & \quad 00111 & \quad 01100 & \quad 01001 & \quad 11100 & \quad 11000 & \quad 01111 & \quad 1.0 \\
01101 & \quad 11001 & \quad 11101 & \quad 00000 & \quad 00001 & \quad 00100 & \quad 10100 & \quad 0.2
\end{align*}\]

using indicators to mark which variables are in/out of the model. The top ranked models are cubics. The MAP model is a cubic.
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** Suppose our data $y_i, i = 1, 2, \ldots, n$ are jointly independent scalars sampled from a mixture model with $m$ components $N(\mu_j^{(m)}, \sigma_j^{(m)^2})$, and mixture weights $w_j^{(m)}, j = 1, 2, \ldots, m$, $w_j > 0$, $\sum_{j=1}^m w_j = 1$.

The observation model we consider for the iid $y_i, i = 1, 2, \ldots, n$ is the mixture

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

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, \( \alpha = 1 \) (so uniform, but 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 \(m \sim \text{Poisson}(\lambda)\) with \(\lambda = 10\), which is centred at 10, and tails off above about 20 clusters.
Posterior: The posterior for the model and parameters

\[ \theta(\text{m}) = (\mu(\text{m}), \sigma(\text{m}), w(\text{m})) \]

is then

\[ \pi(\theta(\text{m}), \text{m}|y) \propto L(\mu(\text{m}), \sigma(\text{m}), w(\text{m}), \text{m}; y) \]
\[ \times \text{Dirichlet}(w(\text{m}); \alpha_1 \text{m}) \]
\[ \times \prod_{j=1}^{m} N(\mu_j(\text{m}); 20, 10) \text{Gamma}(\sigma_j(\text{m}); 1.5, 0.5). \]
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 a move $\text{move} \sim U\{1, 2, \ldots, 5\}$.

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

Step 2Ia Simulate $\mu'_{m+1}, \sigma'_{m+1} \sim g(\mu'_{m+1}, \sigma'_{m+1})$. We will take $g()$ 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

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

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 $\text{move} = 2$ 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

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

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.
Summarising the Posterior over models and parameters for the 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 joint posterior of the mean values and weights in each sampled state (i.e., the points are the pairs $\mu_i(t), w_i(t), i = 1, 2, ..., m(t), t = 1, 2, ..., T$. Points are colored by the number of clusters $m(t)$ in the state.

The top figure shows an estimate of posterior predictive distribution $p(y'|y)$ (black line) obtained by averaging the likelihood $L(\mu, \sigma, w, m, y')$ over the sampled states, at each point $y'$ on the $x$-axis), and the posterior predictive distribution $p(y'|y, m)$ 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.