Lecture 11: Reversible Jump MCMC (case study cont). Loss and utility.

Notes and Problem sheets are available at

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

and via the MSc weblearn pages.
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 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 $(\mu(t), \sigma(t), \omega(t), m(t)), t = 1, 2, ..., T$ 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 (ie, 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.
Decision Theory in the context of statistical inference

For action $\delta$ and truth $\theta \in \Omega$ our loss function is $L(\theta, \delta)$.

For data $y \sim p(y|\theta)$ $y \in \mathcal{Y}$ and estimator $\hat{\theta} = \delta(y)$, $\delta : \mathcal{Y} \to \Omega$ is an action for each $y \in \mathcal{Y}$, with risk

$$R(\theta, \delta) = \int_{\mathcal{Y}} L(\theta, \delta(y))p(y|\theta)dy.$$ 

If we have a prior $\pi(\theta)$, posterior $\pi(\theta|y)$ and marginal likelihood $m(y)$ the posterior expected loss is

$$\rho(\pi, \delta|y) = \int_{\Omega} L(\theta, \delta)\pi(\theta|y)d\theta.$$
The prior average risk \( \rho(\pi, \delta) = E_\theta(R(\theta, \delta)) \) is closely related, 
\[
\rho(\pi, \delta) = \int_\Omega \int_\mathcal{Y} L(\theta, \delta(y))p(y|\theta)\pi(\theta)d\theta dy,
\]
since \( \rho(\pi, \delta) \) is equivalently 
\[
\rho(\pi, \delta) = \int_\mathcal{Y} \rho(\pi, \delta(y)|y)m(y)dy.
\]
\( \rho(\pi, \delta) \) determines a total ordering on \( \delta \), whilst \( R(\theta, \delta) \) may give a different ordering for each \( \theta \). The Bayes estimator \( \delta^\pi \) for \( \theta \), 
\[
\delta^\pi = \arg\min_\delta \rho(\pi, \delta)
\]
is given for every \( y \in \mathcal{Y} \) by 
\[
\delta^\pi(y) = \arg\min_\delta \rho(\pi, \delta|y).
\]
Admissibility

If we accept the loss function $L(\theta, \delta)$ we would never use an estimator $\delta_0$ which was “never better and often worse”. If there exists an estimator $\delta_1$ satisfying

$$R(\theta, \delta_0) \geq R(\theta, \delta_1)$$

and for at least one $\theta_0$,

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1)$$

then we say $\delta_0$ is not admissible. Otherwise it is admissible.

Estimators that seem reasonable (recall James-Stein beats MLE) need not be admissible.
Proposition 2.4.22 (CR-TBC): If prior $\pi$ is strictly positive on $\Omega$ with finite Bayes risk, and the risk, $R(\theta, \delta)$, is a continuous function of $\theta$, then Bayes estimator $\delta^\pi$ is admissible.

Proof: Suppose the opposite. For some $\delta'$, $R(\theta, \delta^\pi) \geq R(\theta, \delta')$ for each $\theta$, and there exists $\theta'$ and an open neighborhood $C'$ of $\theta'$ such that $R(\theta, \delta^\pi) > R(\theta, \delta')$ for $\theta \in C'$. Integrating both sides of the inequality,

$$\int_{\Omega} R(\theta, \delta^\pi) \pi(\theta) d\theta > \int_{\Omega} R(\theta, \delta') \pi(\theta) d\theta,$$

that is

$$r(\pi, \delta^\pi) > r(\pi, \delta').$$

But that is impossible as

$$\delta^\pi = \arg \min_{\delta} r(\pi, \delta)$$

by definition. [EOP]
Actions, rewards, utility and loss: Let $r(\theta, \delta)$ denote the reward if our action is $\delta$ and the truth is $\theta$. We assume bounded rewards, $r_{\min} \leq r \leq r_{\max}$. Let $U(r)$ denote the utility of reward $r$.

Utility is the opposite of loss. In terms of our previous notation, 

$$L(\theta, \delta) = -U(r(\theta, \delta))$$

We replaced one function $L$ by two, $U$ and $r$. Since $\theta \sim \pi$ and $y \sim p(y|\theta)$, $R = r(\theta, \delta)$ is a random variable. Let $P(r)$ denote the distribution over rewards determined by our model.

Till now we chose the action minimising the posterior expected loss. Now we maximise the expected utility $E(U(R))$. Same thing. The utility function (or loss function), must be elicited.
Example: urn with 100 balls, red/black. Choose a color $\delta$, say black. A ball with color $\theta$ is drawn from the urn. If $\theta = \delta$ (if it is black) we get £1. Let $\phi = \Pr(\theta = \text{black})$ (ie, the truth), and let $\pi(\phi)$ be our prior for $\phi$.

Possible rewards are $r = 0$ and $r = 1$. Suppose $U(0) = 0$ and $U(1) = u$ with $u > 0$. Then

$$P(r = 1) = E_{\phi}(E(I_{\theta=\text{black}}|\phi))$$

so $P(r = 1) = E(\phi)$. The expected utility of choosing black is

$$E(U(r(\theta, \text{black}))) = P(0)U(0) + P(1)U(1) = uE(\phi)$$

and

$$E(U(r(\theta, \text{red}))) = uE(1 - \phi).$$

Choose black if $E(\phi) > 1/2$ (and indifferent if $E(\phi) = 1/2$).
Choice of reward distributions Given a choice between two reward distributions $P(r)$ and $P'(r)$, I prefer the one maximising the expected utility

$$P \succeq P' \iff E_{P}(U(r)) \geq E_{P'}(U(r)).$$

Represent preference relation using $\succeq$, $\preceq$ and $=.$

Example: As before but two urns. If $\phi_i$ is proportion black in urn $i$ and $\pi_i(\phi_i)$ prior for $\phi_i$, $i = 1, 2$ then $P_i(1) = E_{\pi_i}(\phi_i)$ and EU’s are $uP_i(1)$. Choose reward distribution/urn with larger EU.

We may start with a preference over reward distributions. The utility function we elicit must reproduce that relation. The expected utility hypothesis says a utility function representing our preference over reward distributions exists.
Jensen’s inequality: If $f(x)$ is concave and $E(X), E(f(X)) < \infty$

$$E(f(X)) \leq f(E(X)).$$

Proof: $f : \mathbb{R} \rightarrow \mathbb{R}$ is concave if for all $x, x_0 \in \mathbb{R}$,

$$f(x) \leq f(x_0) + (x - x_0)f'(x_0)$$

ie, graph below tangent. Since this holds for all $x, x_0 \in \mathbb{R}$, take $x = X$ and $x_0 = E(X)$ and expectations on both sides.

Remark: If $f(x) < f(x_0) + (x - x_0)f'(x_0), x \neq x_0$ (so that $f(x)$ is strictly concave) and $X$ is not constant then

$$E(f(X)) < f(E(X)).$$

Example: in a choice between an average reward $r_0 = E(r)$ and a random reward $r \sim P(r)$, for a concave utility we take the average reward, since $E(U(r)) < U(E(r))$ so $E(U(r)) < U(r_0)$. 